The artsy corporation

Table of Contents

Introduction In the Arts Corporation Case, we are presented with a lawsuit put together by all female employees that work at Artsy Corporation.

With this lawsuit the female part of the company tries to make a point that they have been discriminated in the workplace due to their gender. They make a statement that gender would affect certain factors such as: pay, hiring, promotions and other elements which are involved with the career of Artsy’s employees. This paper will put together statistical information and analysis based on the data we retrieve in order to give Artsy’s lawyers a perspective of what the company truly looks like in order for them to be able to put together a defense.

The data put together for us to analyze is based on the payroll of Artsy’s 256 employees at one of their facilities. This specific data was selected by usage

of simple random sampling (the sample represents the population, i. e. every employee’s pay rate and working conditions). The data includes: •an identification number (IDNUMBER) that would permit us to identify the person by name or social security number, •the person’s sex (SEX) where a 0 denotes female and a 1 denotes a male, •the person’s job grade in 1986 (GRADE), hierarchy level at company the length of time (in years) the person had been in that job grade as of 12/31/86 (TING), and •The person’s weekly pay rate as of 12/31/86 (RATE). (this being the most important point of concern) In order to analyze this data statistically we will create multiple regression models. In this case, we will consider Pay Rate as a dependent variable and Gender, Job Grade and Time in Grade as independent variables. We will work through the data and try to find evidence to show whether or not gender plays a role in the company by influencing the salaries.

On the other hand, if we find out that gender does not play a role in influencing the salary then we will try to find the other independent variable that does make a difference. ? Part 1: Descriptive Statistics Defining Important Terms In this paper we will be making use of descriptive statistics which is basically a way of collecting, summarizing, and analyzing our data in order to come to a conclusion. This is very valuable for us since it will allow us to analyze of large group of numbers such as the set of data we are considering.

When analyzing each specific set of data we will be considering both the central tendency and variation of the specific variable we are discussing. Also in this paper we will refer to the “ mean” in specific parts, where the word basically refers to a measure of central tendency, it is the arithmetic

average of the numbers in our data. In order for our reader to have a better understanding of mean and arithmetic average, we have given an example below: By using our data and the software available to us, we were able to conclude that the average pay rate of females in Artsy Corporation was $833 er week, whereas the average pay rate of males in Artsy Corporation was $1128 per week. With this information we can have an idea that males have a higher salary (This is a quick example which does not take the other variables into consideration). The last important term that is necessary to clarify in order for a better understanding is: standard deviation. The standard deviation is a number that measure the spread of the data in relation to the mean number. It gives us the scatter of the information in terms of percentage, giving us an idea of how close together or separate our full data set is.

Data Analysis of Pay Rate Now that we have a better understanding of the concepts we can start analyzing our data table. Our first variable to analyze is pay rate, since it is the one of our concern. The lowest pay rate is $579 per week, and our highest is $1552 per week, giving us an average pay rate of $931 per week. Our first standard deviation is $229 per week. Basically what this means is that 68% of our data (1st standard deviation) is either $229 above or below our mean of $931. Below we have put together a box-whisker plot of the pay rate in order to have a better idea of the salaries at this corporation: Figure 1

In the plot above we are able to come to various conclusions. The first points we have to analyze are our quartiles: •25% of the employees have a weekly pay rate less than $762. •50% of the employees have a weekly pay rate

between $762 and $1073. •25% of the employees have a weekly pay rate above $1073 •Our middle number here is $865, which means that 50% of our employees make more than this, and 50% make less •The little red box on the right of our data is an extreme outlier which basically signifies a point which is out of the context compared to other numbers.

We assume that this number is the pay rate of the executive or manager of the branch which is much higher than those of the employees, therefore it should not be used to compare to others. By simply looking at this graph we are able to see that the lower 25% of the salaries are much more agglomerated at the end and seem to all be in the same range. However, the salaries of those in the highest 25% of the data are much more spread out since the line is much longer at this end. The lines at the end of the box give a sense of how close together, or spread out our information is.

In this case the higher salaries are more spread out. Data Analysis of Pay Rate vs. Gender In the introduction we talked about the average salaries of men and women. We discussed how they were very much different even though we still had not taken into consideration the other variables given to us in the data set. Below we have created a box and whisker plot where the pay rate of the men and women of the company are put side by side in order to have a better idea of how they differ. Figure 2 From the data displayed in the plots above we can see that there is a big difference between male and female salaries.

The bottom 25% of the male employees make approximately the same amount as the women in the intermediate portion (from 25%-75%, middle

50%). Since this is a sample which we will base our whole population on, then we can say that this information pertaining to this sample represents our company’s numbers in general. Also, once again there are red squares to the right of our plots. In this case there are 4 red boxes (2 are overlapping showing a darker tone of red), which show that the higher paid females in the company are seen as outliers when analyzing them statistically.

Women that have a higher pay rate are not in the normal patterns of Artsy and therefore are statistical outliers compared to the rest of the company. Once again, since this is simply an initial data observation without taking into consideration our other variables we cannot make a conclusion, however; we can see that there is a tendency of higher salaries towards men, and we can infer that there is a possibility that gender does play a role in pay rates. Data Analysis of Grade We can now analyze the grade of our employees and how this affects the salary.

The grade is simply the hierarchical position of our employees in the company where they are scaled from 1 to 8 (1 being the lowest, and 8 being the highest). Below is a table of each of our employees in terms of grade. Figure 3 Looking at the graph, we can see that there is a similar distribution of employees at both the top management positions, and also the lower position. It seems to be very well distributed which takes away the idea that the higher the grade the fewer employees there are. Figure 4

Above in Figure 4 we have put together a scatter plot of our employees pay rate and how it differs as they go up in their grade levels. Each circle represents one employee and their position symbolizes what level they are

at, and what pay rate they are given. Also we have made a gender code where we are able to see the difference between males and females. The red box (1) represents male employees, while the blue circle (0) represents the female employees. There is clearly a trend in this graph where the higher an employee is based on his grade level, the higher his salary will be.

By simply observing the graph, we can see that on average, men reside in the higher grade levels with higher pay rates. On the other hand the women reside on the lower grade levels with lower pay rates. Also we can even see that when men and women are together on the same level the men tend to be on the higher side of the pay rate per grade, while women are closer to the bottom proportion of each grade level. To have an idea in relation to the numbers, in grade level 2 all of the occupants are female employees. However, when taking into consideration grade level 7 only 34% of the employees here are female.

All of this information from this specific graph shows further evidence that there is gender discrimination within Artsy Corporation; however we will continue to seek more information to have statistical proof. Data Analysis of Time within Grade Our last variable which we will discuss is Time within Grade. This variable simply shows how long each employee has stayed within his or her grade level in terms of years. On the precious page we saw a graph which shows the pay rate in terms of time within grade level with the gender distinction in order to observe differences.

Each dot represents an employee, whereas male is symbolized by the red square (1), and the female is symbolized by the blue circle (0). We can see

that, on average, women have held their grade for a lesser time when compared to men’s time in grade. In addition to that, we can also see that women and men who have the same amount of time in grade results in women earning a lower pay rate than men. For example, a male who has held a grade for 0. 5 years has a pay rate of $1413 per week as opposed to a female who has held a grade for 0. 5 years has a pay rate of $605 per week.

However, the source of this large difference can most likely be because of a difference in grades. Further into the report, we will explore if there is a strong relationship between time in grade and pay rate or not. By using descriptive statistics, we found that there are signs that point to pay discrimination based on gender. However, this is not sufficient evidence to prove that there is, in fact, job discrimination. Therefore we will move onto the next part of our paper which will analyze our data in terms of regression.

PART 2: Regression Analysis

Now the second part of the paper will commence where we will be analyzing our data in terms of regression. This regression part will help us analyze with more depth how pay rate alters in relation to our independent variables (gender, grade and/or time in grade). The use of regression will be essential to find out if there is a gender discrimination affecting the pay rate. To start off the regression section we will need to choose a significance level which will be our percentage of error. For this paper we chose a significance level of 1% which means that we will have close to zero amount of error in our model.

The choice of this 1% is basically because since we are being sued by employees and we are providing our lawyers with the best information available trying to avoid any possible error. This will definitely provide them with a solid defense for Artsy Corporation. In regression we need to choose one hypothesis which we ultimately will be testing. Hypothesis testing basically means setting up two opposing statements which are called the null hypothesis and the alternative hypothesis. Both statements cannot be true since they are mutually exclusive.

Below we have displayed both our null and alternative hypothesis: ? Ho (null hypothesis): There is a linear relationship between our employees Pay Rate and the independent variables which we have defined (gender, grade, time in grade). We believe this to be true until we are proven otherwise. ? H1 (alternative hypothesis): There is no linear relationship between our employees Pay Rate and the independent variables which we have defined (gender, grade, time in grade). We reject the Ho (null) if the p-value of any of our variables shows up larger than our significance level (1%).

The P-value is the probability that the sample data would occur if a pre-defined null hypothesis (H0) were in fact true in the population. We use each of the individual p-values to compare to our significance level (1% which we chose above). If the p-value is less than 1%, then, we reject the H0, if not, we do not reject. If the p-value is over the significance level, we fail to reject the null hypothesis, because as statisticians the probability that our data sample would occur is not statistically significant.

Statistically significant means, that maybe the data occurred only due to chance alone, rather than being due to other independent variables. With use of software we were able to come to an equation which predicts the pay rate based on our independent variables (gender, grade, time in grade). This equation is displayed below: Pay Rate = 527 + 59. 6 Gender Coded + 30. 8 Time In Grade + 75. 0 Grade R2= 82. 3% S. E (Standard Error of the Estimate) =$97. 0601 per week Now it is critical to understand the equation above and how it is structured the way it is.

The 527 comes from our constant which is seen as the starting point of the pay rate. If all other independent variables were to be 0 then the constant (527) would equal the pay rate. The second part of the equation is the gender variable. Gender is a coded variable since it is in qualitative terms. However, since regression only works with quantitative numbers the gender is coded as 0 for female, and 1 for males. The second independent variable is time in grade (years) which is already a quantitative variable so we simply use the number as it is.

The last variable to this equation is the grade level, which in this case is qualitative. We will create 8 new variables (ex: grade 1) for each grade level and give the employee a 1 if they are in the grade level and a 0 if they are nor present in that grade level. Now that we fully understand each of the variables in our equation we can go about explaining how the equation works. As mentioned before the number 527 describes the payment of an employee that has 0 for every other independent variable (a female, 0 years in grade, grade level 0).

We cannot interpret this number since it is not applicable being that grade level 0 does not exist. The second number we will be analyzing is the 59. 6 from the gender variable. This number means that, on average, if the employee is male (1), his pay rate will be $59. 6 per week higher than that of a females pay rate, while keeping all other variables constant. This part of our equation once again shows us signs of gender discrimination since we see that with all other variables equal, me still make $59. 6 per week more than our female employees.

Our third number to analyze is the 30. 8 which is in the Time in Grade independent variable. This number means that, on average for each additional year an employee spends in his or her grade; their pay rate should increase by $30. 8 per week, once again assuming all other variables constant. This shows us the normal flow to companies that the longer employees stay in the job at their grade level their pay rate will increase with their level of experience. Our last number to analyze in the equation is 75 for the Grade level variable.

This means that, on average, for each increase in grade level, the pay rate should increase $75 per week. Once again this shows the traditional hierarchy in companies, where higher level employees receive higher salaries. The R2 of this equation is 82. 3%. The R2 shown above together with the equation tells us how much of the variation in these pay rates are due to the variables which we have just described. This means that in this case, with an R2 value of 82. 3%, 82. 3% of all the differences in pay rate in our set of data are explained with the use of our 3 independent variables (gender, grade, time in grade).

This tells us that statistically we can correctly prove there is a relation between the pay rate and our independent variables 82. 3% of the time. The last part to our statistical information above is the S. E (Standard Error of the Estimate), which is equal to 97. 06. The S. E. displays the variation of our predictions. This number shows how our predictions might fluctuate, basically meaning that by using this equation to predict our employees pay rate we can be off by ±$97. 06 per week. Basically we can have a number that is either over or under the actual number by $97. 6. This tool is very important because it slims down our margin for error. Before in this paper we mentioned that our standard deviation for our pay rate was $229 per week. Now with the regression model we have come to a lower error of $97. 06. We have reduced our error percentages by about 58% with the regression model. Now that we have already created our initial regression model it is necessary for us to create individual regression models for each of the variables in order to find out how much of the variation is caused by each of the independent variables.

Our first individual regression model is using our gender variable. The model that will predict the pay rate with just gender as a predictor variable is: Pay Rate = 833 + 295 Gender Coded R2= 36. 9%S. E = $182. 554/week As we have discussed in our initial model, since gender is either male or female being a qualitative variable, it is displayed as either 1 or 0. In this case we are able to see that on average, a male’s pay rate is $295 per week higher than the pay rate of a female. In the previous regression model this number was much lower ($59.), however, here we are individually selecting variables and isolating them in order to find out how much each of the

variables affect the pay rate. Since our R2 here is 36. 9%, we can say that 36. 9% of the variation in pay rate may be explained simply by knowing the gender. The second individual regression model that we will create is considering time within grade. The model that will predict the pay rate with just ‘ time in grade’ as a predictor is: Pay rate = 788+82. 3 (Time in Grade) R2= 29%S. E=$193. 734/week This model tells us that for every additional year within a grade level the pay rate increases, on average, $82. per week. However, since here we see an R2 which is a low 29%, we can come to the conclusion that the time a person has worked within a grade level is not significant in terms of their pay rate. We believe that this happens because the time within grade does not account for the experience the person has in the company overall. A person in grade 8 might have 10 years in the company, however he might have just been promoted to grade 8 therefore he has a low time within grade. On the other hand there might be someone in grade 1 for the past 5 years.

Obviously their pay rate is cannot be based on time within grade even though it has some affect to the overall model. However, with this percentage of 29% we consider this to be insufficient proof of variation. Our last individual regression model which will be created is considering the grade level. Once again we have made changes to this variable where we separated out the grade level variables in order to create a specific variable for each grade level. We have created 8 variables with a 1 and 0 possibility, 1 being in that grade level and 0 being that they are not in that specific grade level.

Below is our individual grade regression model: Rate = 671 + 694 Grade_8 + 501 Grade_7 + 385 Grade_6 + 226 Grade_5 + 161 Grade_4 + 161 Grade_3 + 54. 7 Grade_2 R2= 81. 8%S=$99. 3584/week Above we have displayed how the pay rate increases in each grade with respect to grade level 1(the initial grade level at the company). What this basically means is that grade level 8 employees on average make $694 per week more than grade level 1 employees. At the same time grade level 4 employees make $161 per week more than grade level 1 employees.

All the grades in the model above are being compared to the lowest grade level possible (grade level 1) since it is not on the regression model, and we can relate all numbers back to it. Now in terms of the R2, we see that there is a very large number in comparison to all the other individual models. Basically 81. 8% of the pay rate can be explained simply by knowing the employee's grade level. The lawyer's defending Artsy can definitely use this information since they can say that 82% of the time the pay rate is defined due to the employee's level inside the company.

It can also be compared to the individual gender model to show how the 82% is much more significant than the 36. 9% we found before. After running the initial regression model together with all of the individual ones, we have come to develop a new regression model where we will use our grades separately as we have shown above in order to see if we increase our R2 and reduce our Standard Error. The new regression model is displayed below: Rate = 632 + 46. 9 Gender Coded + 26. 9 TInGrade + 60. 9 Grade_2 + 151Grade_3 + 168 Grade_4 + 210 Grade_5 + 356 Grade_6 + 434 Grade_7 + 613 Grade_8 R2= 85. % S. E=$89. 29/week Once again this model shows

how the pay rate changes with respect to each individual variable. In the gender variable it tells us that on average men make $46. 9 per week more than women at Artsy Corporation. Also in terms of the grade levels we need to remember that each grade level coefficient is being shown with respect to the lowest grade level at the company (grade level 1). All the ways of interpreting this model will remain the same as the ones which we have shown before. Since this model has shown us a larger R2 of 85. 4% compared to our 82. % from our initial model, we will continue to use the new model since it displays where the variation comes from at a more accurate level. Also another factor that shows us that our new regression model is better from our old one is our S. E. The standard error in our initial model was $97. 06 per week, whereas our new regression model has an S. E of $89. 29 per week. We have reduced our error by $7. 77 or 8%. Once again this is a good sign since it gives the lawyers a more accurate display of information and reducing our error in order for them to have a better more solid defense when making the case against gender discrimination.

Before we can start using our regression model we need to check and see if all of our variables can be present in our model. In order for variables to be accepted in regression models they need to be in two specific conditions: linearity and equal variance. We will check these two conditions using a normal plot of residuals and residuals versus fits plot. We would first want you to understand what linearity and equal variance means. Equal variance means that the variability in pay rates is the same regardless of the independent variables having either high or low values.

Linearity is to see when the pay rates vary directly in proportion to our independent variables (gender, time within grade and grade level). Also the residual is the difference between the observed value of the sample and the predicted values of the pay rates for a given independent variable. Now that we have explained both of the test in order to prove that these individual variables should be present in our final regression model we shall show each of the tests (linearity and equal variance) for each of the variables in consideration.

On the next page we have displayed the first scatter plot, which shows the pay rate in relation to gender: Correlation: 0. 608 On the scatter plot above each circle represents one employee at the company. All of the dots above the number 0 account for the female employees at the company, and those circles above the number 1 are for the male employees of the company. Since there are only two possibilities it makes it difficult to see the linearity, however, we are still able to see a pattern where male employees have a higher pay rate them female employees.

There is a slight increase present in the scatter plot above. Now we will check to see if this variable passes on our second test which is our equal variance test: Each point on the graph represents a residual. In order to determine if there is equal variance within the variable we need to look at both stacks and determine whether they are similar in size or not. Equal Variance happens when the data stacks are roughly the same size. There is unequal variance when one data stack is more than twice the size of the other.

Our next variable to analyze will be time within grade and we will check to see its linearity and equal variance. The graphs are displayed below: In the scatterplot above, even though there seems to be no trend at all with various points spread out, there is a trend which shows that the longer each employee spends within the grade, the higher their pay rate will be. We can therefore conclude that this variable passes the linearity assumption. Below we have created the equal variance assumption scatter plot in order to see if this variable passes this assumption as well.

Correlation: 0. 538 Once again since our data stacked in the scatter plot above seems to be of the same size we assume again that our equal variance assumption is not violated. This makes this variable accepted to run the regression model just as the prior did. Finally our last variable that we need to test is our grade variable. We need to check just as we did with the two previous variables, whether or not they would be acceptable for our regression models by looking at the linearity assumption and the equal variance test.

Below is our linearity assumption test: Correlation: 0. 876 The test above once again shows us that there is a linear relationship between the rate and grade and therefore we can assume that this variable also passes the linearity assumption test. Below is the equal variance graph which will test to see if it passes the test assumption: Above we see the last graph pertaining to equal variance where we see that all of the stacks seem to have the same size, and even though some are larger than others there is no significant difference to which we should comment on.

It also passed both the linearity and equal variance test. Conclusion After carefully analyzing all of our data and the variables given to us we can come to the conclusion that there are many signs which point to gender discrimination in the workplace due to the first section of our paper. However, when analyzing the regression models in the second part we were able to see that the major contributing factor to the pay rate was the grade level due to its high R2 level of 81. 8%.

Also if the lawyers need to use one of the regression models in order to prove their point we advise they use the fully constructed final model which has the lowest percentage of error, and also the highest correlation between the pay rate and the variables given. Glossary: Regression Model: Statistical technique that uses two or more numerical independent variables to predict the value of a numerical dependent variable. For example, we are using factors such as a person’s gender; position in the company and length of time in the position (all are independent factors) to predict the pay rate (dependent factor) of an employee.

Central Tendency: It is the extent to which the values group around a central value. Variation: It is the amount of scattering of values away from a central value. Arithmetic Average: It is the sum of a series of numbers divided by the count of that series of numbers. Linear: If rate of change of Pay Rates is constant or not. Before attempting to make any regression model, we need to determine if there is a significant relationship between Pay rates and other variables (gender, grade and time in grade). If there is no significant relationship then making a regression model will not be useful or accurate.