

[](https://cidse.engineering.asu.edu/)

**IEE 578: Regression Analysis**

**Instructor : Dr. Douglas C Montgommery**

Final Project : Is the Pay Gap Real ? A Regression Approach

**By:**

**Arvind Rajendran**

**1216546123**

**:: INTRODUCTION ::**

The Pay Gap, one of the questions that plagued and still continues to plague the mind of mid level work force in every industry. But technically what is the pay gap, as defined by the Organisation for Economic Co-operation and Development, the gender pay gap is "the difference between median earnings of men and women relative to median earnings of men."

Per the most recent data from the United States Census Bureau, women on average earned 82 cents for every $1 earned on average by men in 2020. Women of color, regardless of education, are often channeled to lower-paying jobs compared to White women operating at a similar skill level.

The wage gap between men and women has a long history, with some twists many people may not realize. But that is out of the scope of this report. The question I want answered through this analysis are the following :

* Is A Person Paid More Because of Factors/Biases Outside of Their Performance at Work ?.
* Does a white caucasian male get paid a more satisfactory compensation than an African American women who does the same exact job ?.

And many more relatable questions can be chained along with the above questions but that is for another time.

**:: DATA DESCRIPTION ::**

The data used for creating the model was obtained from kaggle.com under HR Datasets from the IBM corporation. The data initially contained 34 columns i.e. regressors which could be used to predict our response which in this case is the payrate of an individual converted into dollar ($) /(per) Hour. The data set contains 311 data points for the analysis.a vague description of the data points are given below along with their encoding values.

There were some columns which can be completely ignored for the analysis such as DaysLateLast30, LastPerformanceReview Date etc., but there are some derivable values from the data given also especially from DateOfHire, DOB ( DateOfBirth), from the two columns we can get the years the employee has spent at the company and the age of the employee, which could be an important input while trying to predict pay rate. The reduced set of regressors are as below:

|  |  |
| --- | --- |
| **Feature** | **Value Range/Description** |
| MaritalStatusID | 0 - Single, 1 - Married, 2 - Divorced, 3 - Separated, 4 - Widowed |
| EmpStatusID | 1 - Active, 2- Future Start, 3 - Leave of Absense, 4 - Terminated, 5 - Voiluntarily Left |
| DeptID | 1- Admin Offices ,2 - Executive, 3 - IT/IS, 4 - Software Engineering, 5 - Production, 6 - Sales |
| PerfScoreID | 1 - PIP , 2- Needs Improvement, 3 - Fully Meets, 4 - Exceeds |
| PositionID | Position ID of employee according to title. |
| State | Office Location according to State |
| Age | Age of Employee |
| GenderID | Male (1) or Female (0) |
| CitizenDesc | 1 - US Citizen, 2 - Eligible NonCitizen, 3 - Non-Citizen |
| RaceDesc | 1- Black or African American, 2- White, 3 - Asian, 4 -American Indian or Alaska Native, 5 - Two or More Races, 6 - Hispanic |
| YearsAtCompany | Time Spent at Company In Years |
| ManagerID | ID of Manager |
| EngagementSurvey | Average of Scores Assigned by peers about engagement in company functioning |
| EmpSatisfaction | 1-5 rating of employee satisfaction. |
| SpecialProjectsCount | No. of Special Projects Done |

**:: EXPLORATORY DATA ANALYSIS ::**

Before the start of fitting any model, the data has to go through a stage of preparation and cleaning, imputation or deletion and encoding variables. This includes finding whether outliers within the dataset contribute to the analysis or if they were just recorded wrong, another part of the data cleaning phase is searching for missing values within the data set.

After finding out the missing points and removing the outliers we have 302 rows of data and 15 possible regressors.

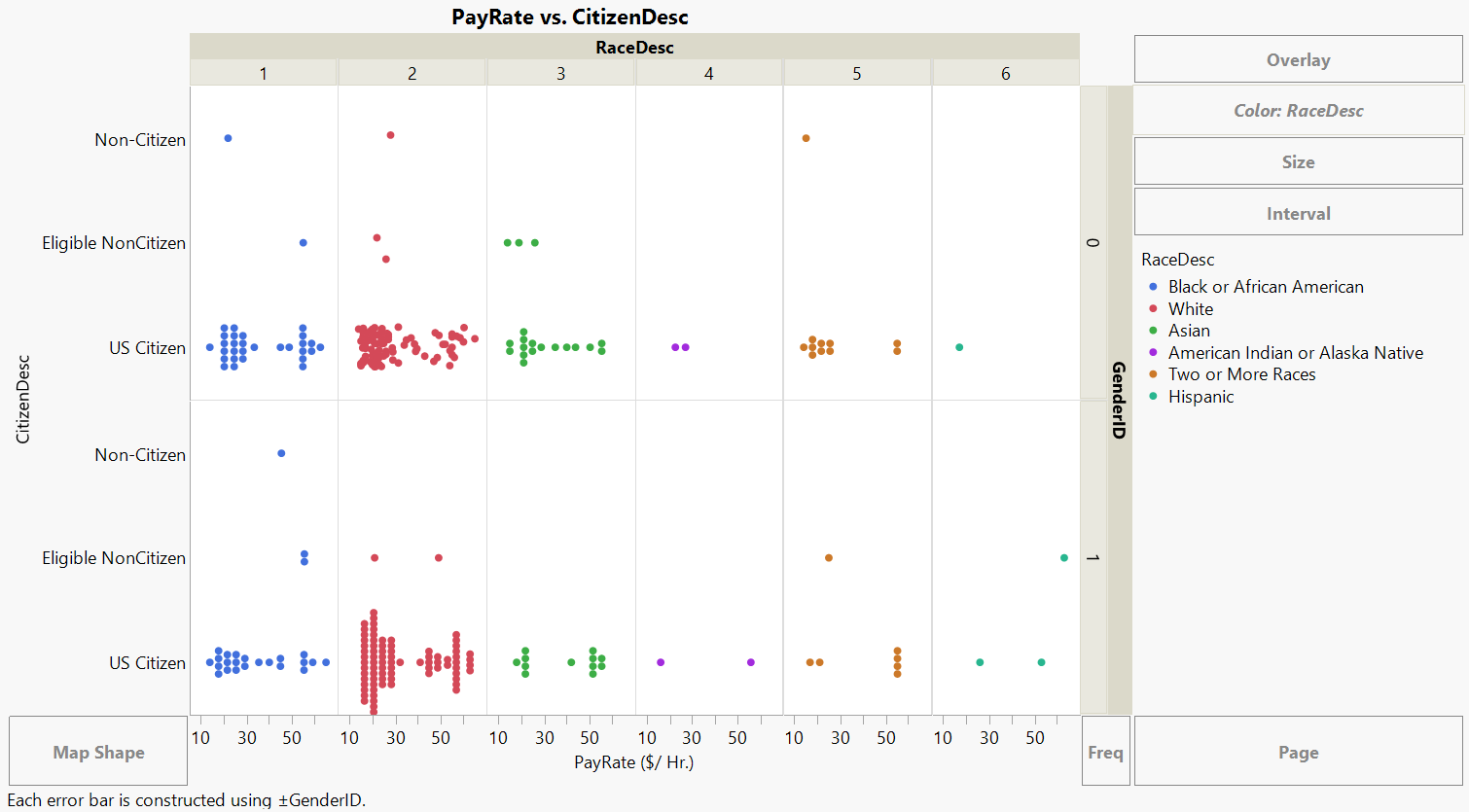
The founder of the company who occupies positon ID 16 and has the highest payrate of 80 was found to be a outlier as her time spent at the company was lower comapratively but her payrate refelected her position and hence has been removed from the analysis.

Among the dataset are also people who work under no manager, these could be an error in recording the data as they have mangaer names but not a code hence these 8 values have also been removed from the model.

A simple graph can tell us loads about the company and their work force. The graph tells us about the distrubution of employees around the different socio econimic divides within the workforce.

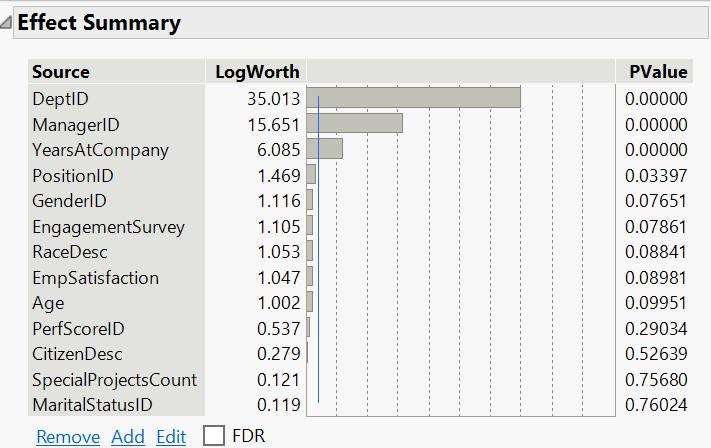
The graph shows that the count of the White(RED) demographic within the workforce can likely skew the analysis in unpredictable ways, especially when we are conducting a regression analysis to check for presence of racial. But as far as the gender population of the workforce goes it looks pretty balanced

GenderID : Female (0), Male (1)



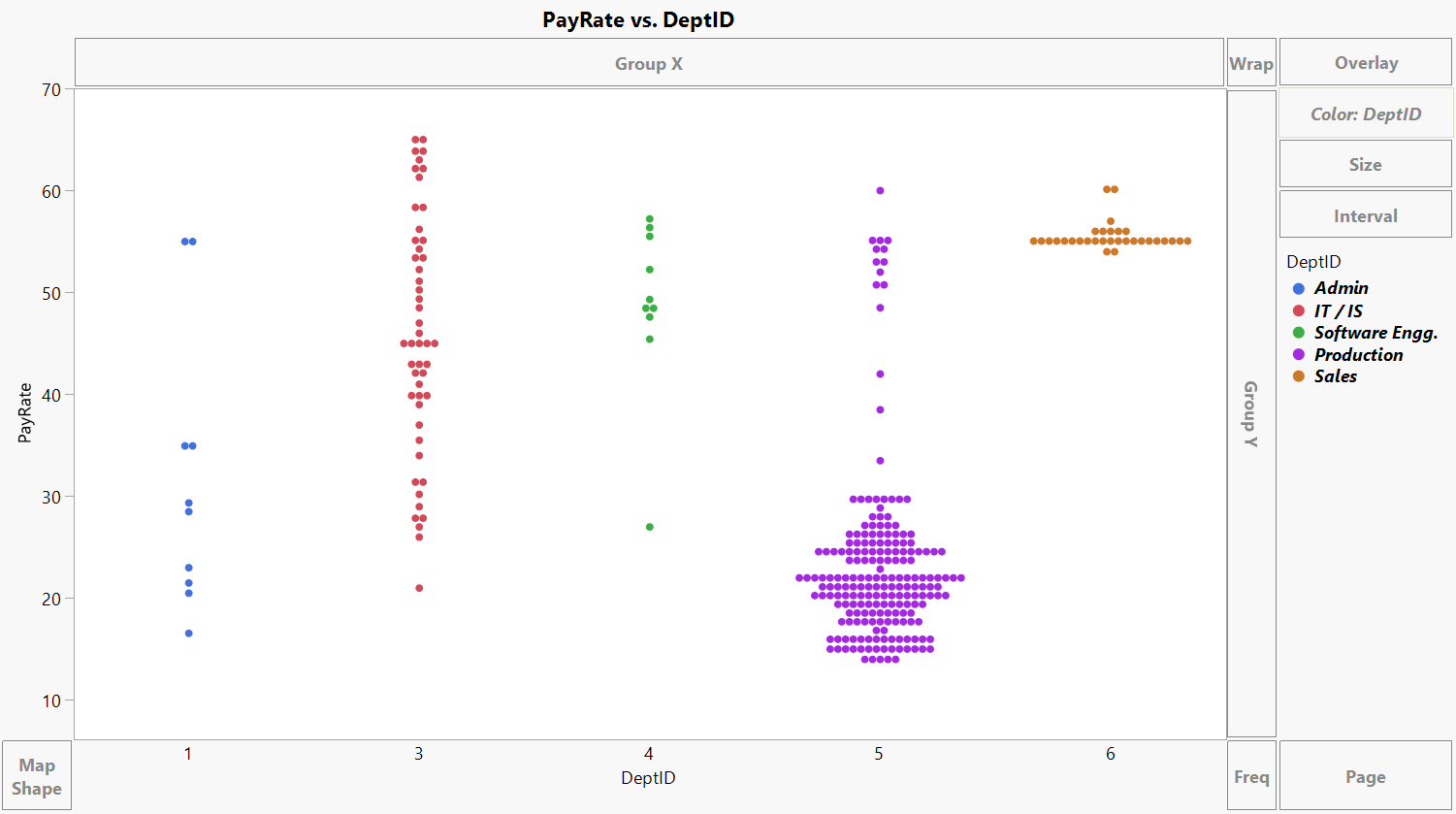
**:: REGRESSION ANALYSIS ::**

We start with a simple regression model which uses all the above regressors and see what the initial analysis has to say about the data and the underlying model. And right of the bat we see :



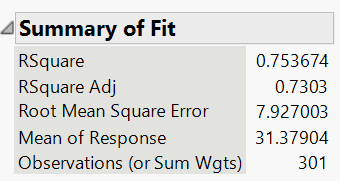
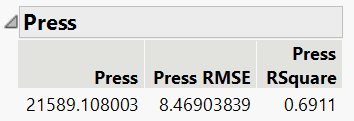
The Dept ID, Manager ID, Years Spent at the Company have higher significance in the analysis which confirms to common logic such as,

* Departments like Sales and IT/IS have the highest paying salaries relative to departments like production where the mean salary is pretty low which can be seen below :

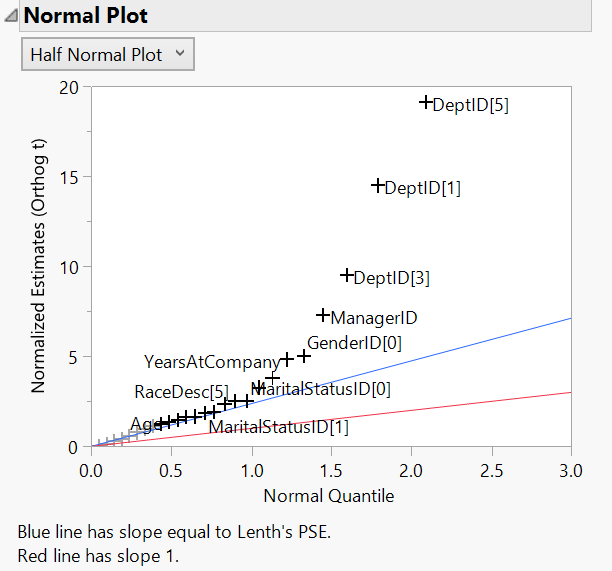


* ManagerID also makes a lot of sense as managers usually get more than those who work under them and managers too have levels where the manager of the highest level would be the CEO of a organization, i.e the manager of production would have a higher paid manager from the admin department.
* Years Spent at a company makes most sense out of the significant factors as a person would have promotions and more compenstaion relative to their stations.

If we look at the R2 values of the fit model we observe nothing special for the model and the PRESS (R2 Predicition) also doesn’t really tell us much about the model and its underlying properties.



To get a more detailed look at the factors we can use the half normal plot :



This plot also gives us the same conclusions as the effect summary which is good. But lets dive into a more complex model such as stepwise regression. Which could tell us some thing important about our approach.

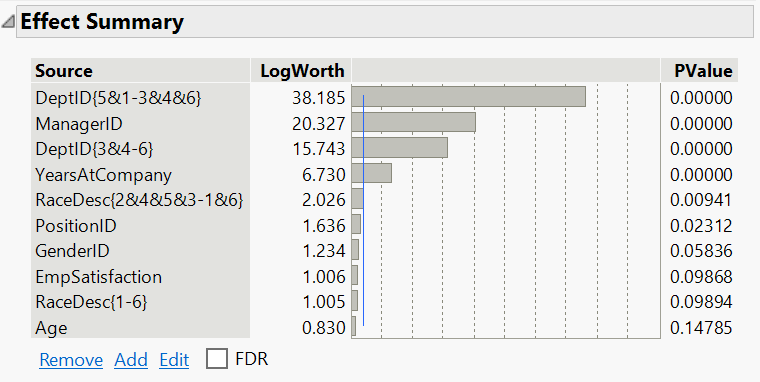
Stepwise Regression :

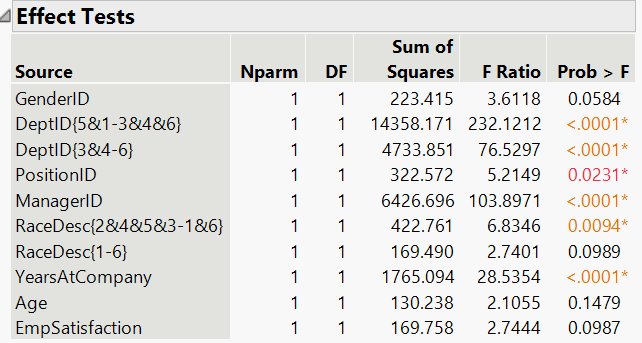
After looking at the Summary of fit data along with the PRESS values we get from running the different types of stepwise regression we can say that there has been a slight increase in R2 values for the three models but the R2 prediction has increase by a lot which means that the elimination of MaritalStatusID, SpecialProjectsCount and CitizenDesc was a step in the right direction, as model has become better at predicting the pay rate.

From the below model though the forward selection algorithm has produced a higher Adjusted R2 value it has almost comparable Prediction R2 we will choose this model to go forward.

|  |  |  |
| --- | --- | --- |
| Backward Elimination | Forward Selection | Mixed Model |
|  |  |  |

Lets look into more details from the forward selection model.



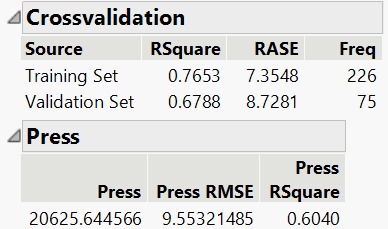


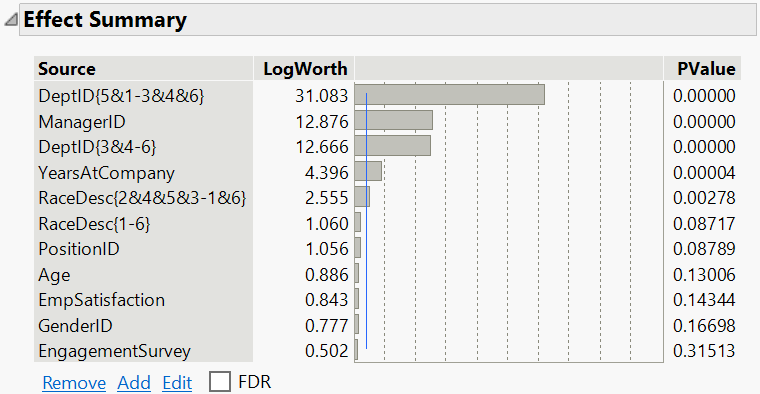
Interesting here we see that RaceDesc, PositionID could be an important factor within the analysis but not as much as the previously declared significant factors Like DeptID, ManagerID & YearsAtCompany.

Now that we have a model lets try to validate the model.

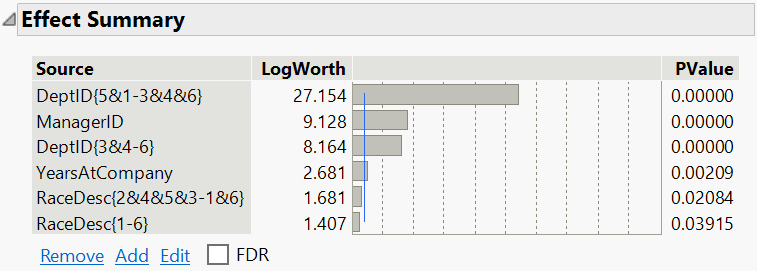
**:: MODEL VALIDATION ::**

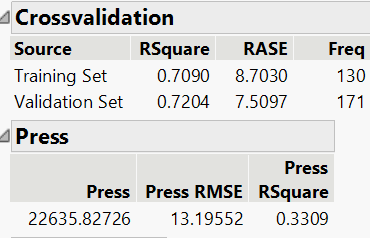
This is an important part of the analysis and here we do two types of validation where in the first validation we will just randomly split the data into training (estimation) and validation (preditcion) data, and in the other validation lets split in a way where the validation set also shows the stratification of the training set, for this we will split it along the RaceDesc column and lets see how the model performs for these 2 situations.





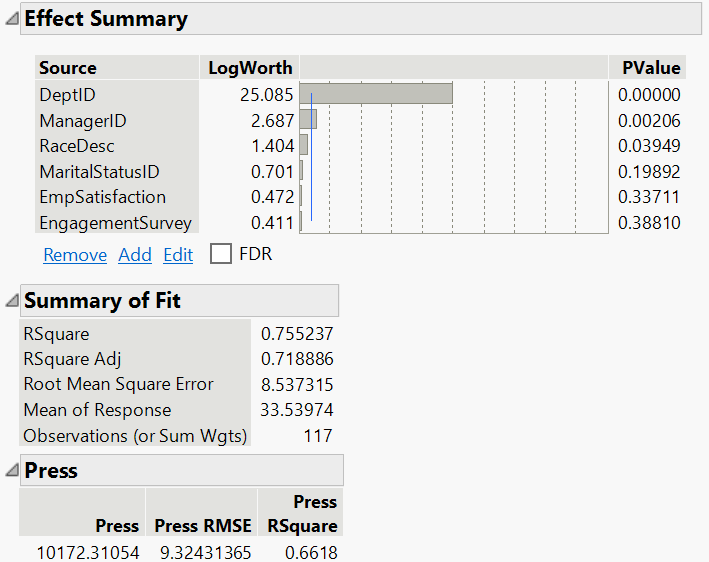
Something very interesting has happened for the first validation step and we can see that RaceDesc according to the model is classified at being a significant factor which affects the PayRate. Interesting so lets dig deeper with our second validation where we split the two sets based on the RaceDesc.



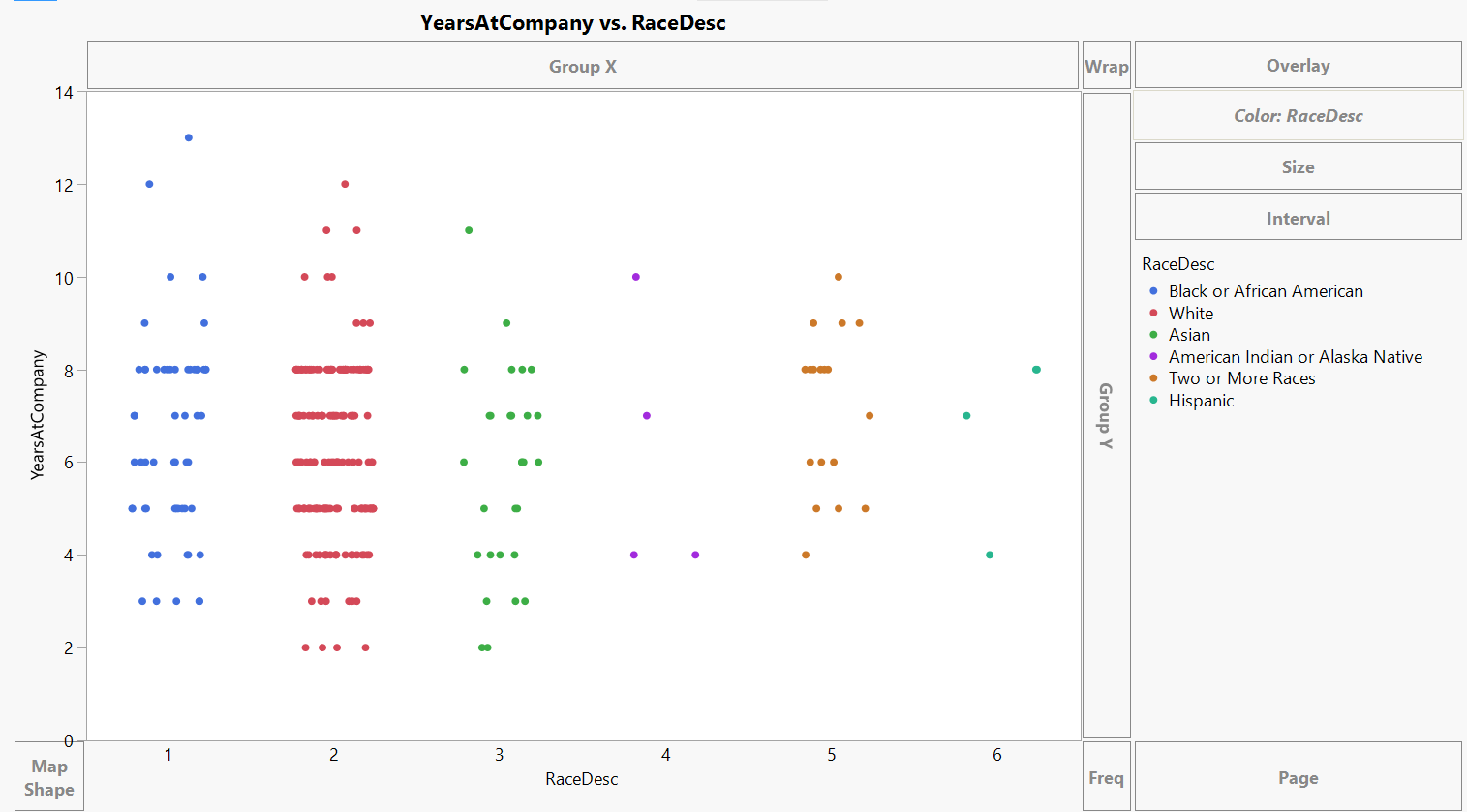


This is very interesting as the model which did so well estimating the basic random split is now struggling with the data set when the validation an training sets are split across the RaceDesc column which gives us an idea about the impoertance of the race of an employee but this was something we knew since before starting the analysis. So what would happen if we create a separate subset where we remove the caucasian(white) community and see how the model changes.

We get more fasicinating results from it:

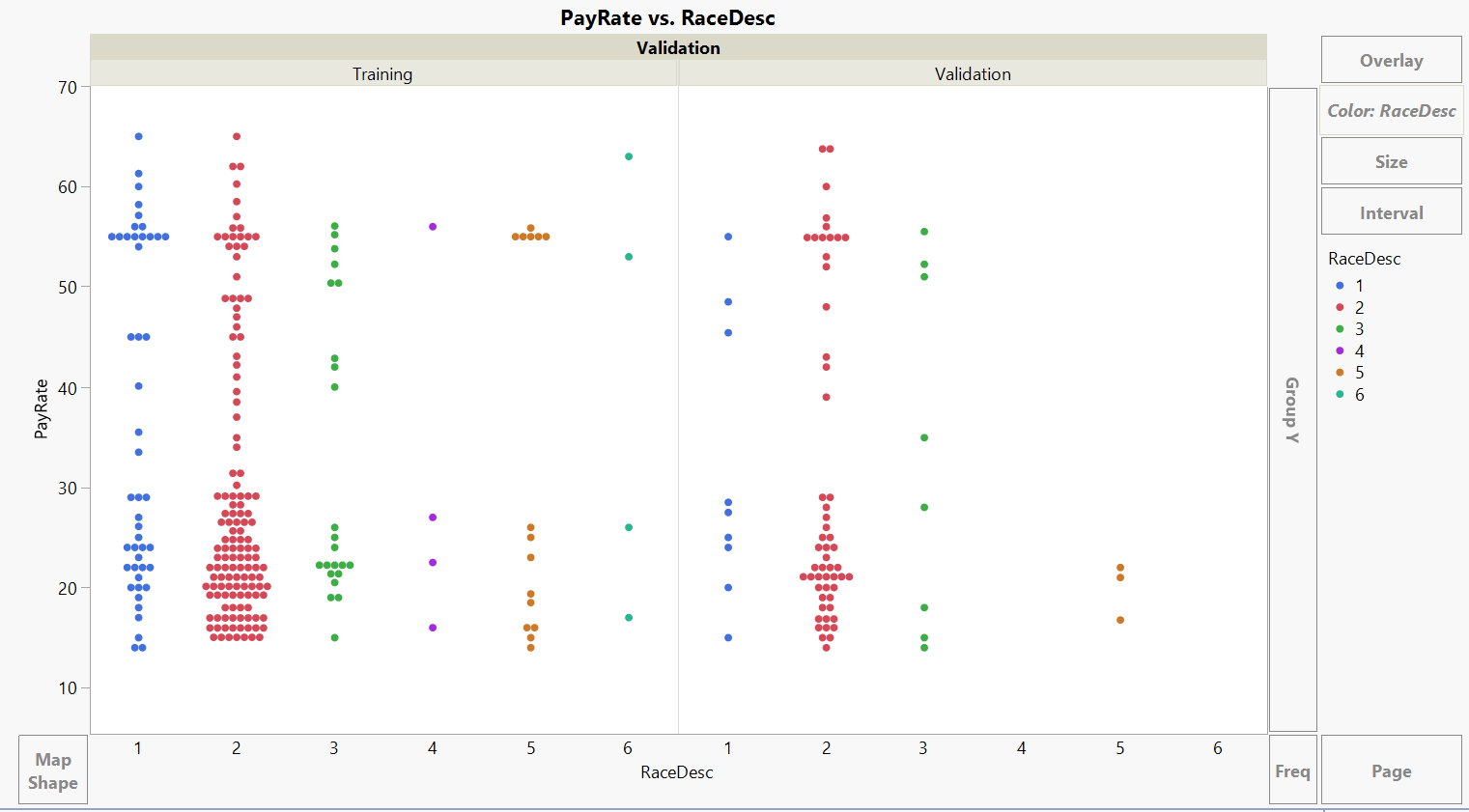


It still tells us that DeptID and ManagerID are very important for determining the payrate but the YearsAtCompany has been dropped from the analysis and this insight can also make a lot of sense when you think about the proportion of people who worked at the company were predominantly White, who skew the analysis in many ways hence the analysis with their category being removed has produced a different asnwer than before.

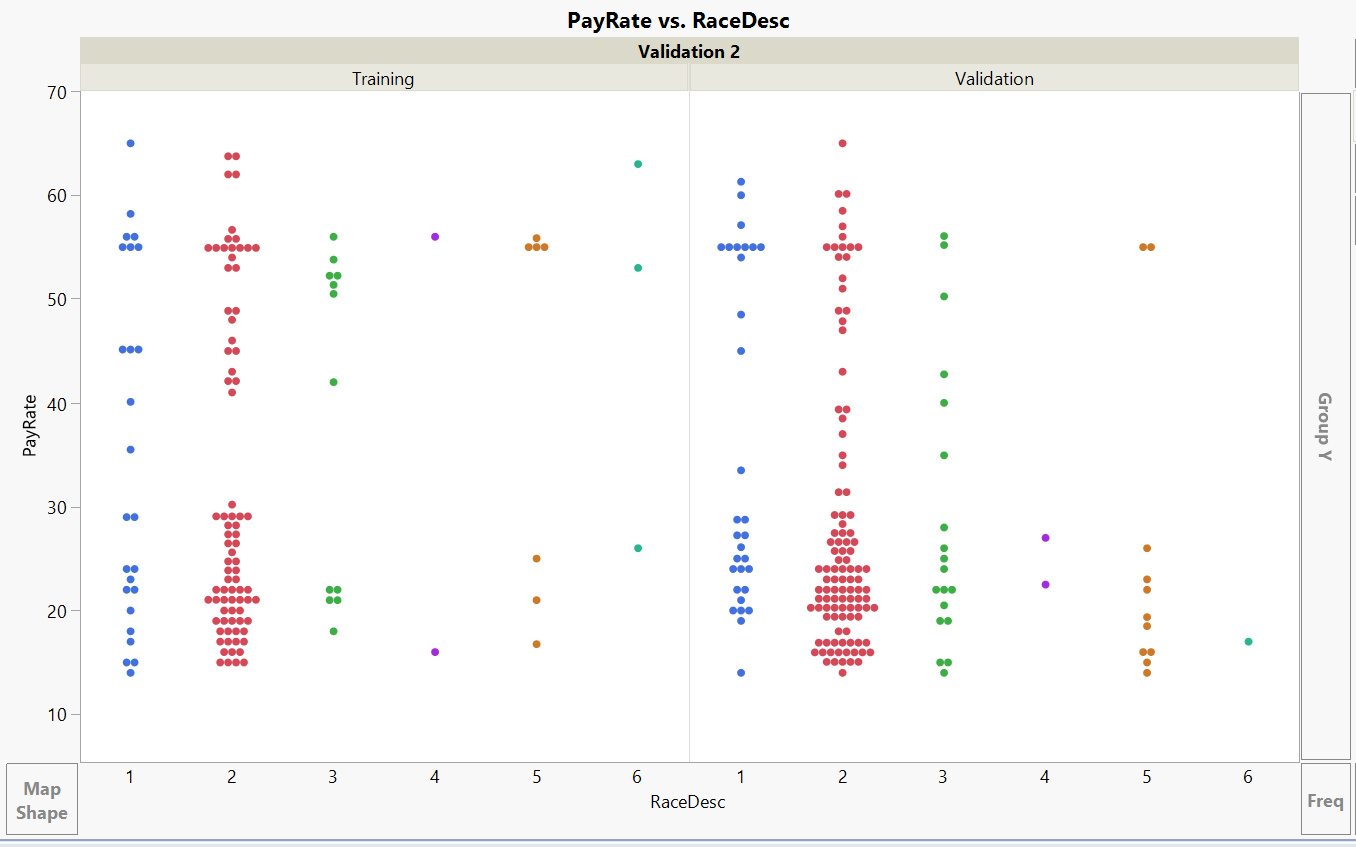


Reasoning for using the two different validation sets are :

Validation 1 : Random



Validation No. 2 : Split on RaceDesc



We can clearly see that the split on the RaceDesc is more generalized i.e. it follows the pattern in the data we have more than the random Validation.

**:: RESULTS & INFERENCES ::**

In conclusion we have done the following steps in this report :

* Data Description
* Exploratory Data Analysis
* Regression Analysis
* Model Validation

And Now we come to the final results of the analysis and we can see that RaceDesc which describes about the race of an employee though we see it pop up in some key areas we cannot declare bias of Race with just this amount of information. To get a clear evidence of this will require more extensive data collection and the data should also be such that we can compare each of the individuals with the same starting point which is out of the scope of this project.

Speaking of bias, throughout our analysis - GenderID, a variable which describes the gender of the employee, never once turned up as a significant factor. Though this doesn’t describe about the wage gap we can say for sure that Gender doesn’t seem to be a criteria used for determining a person salary.

Another interesting part of the analysis was the sheer unimportance of a few features like SpecialProjectsCount, PerfScoreID which should have been features central to predicting a persons payrate but in the analysis these features were shown to be insignificant and after some data exploration we can see that 80% of the people in this analysis have never done a special project for the company, which would also be too demanding of their station. And the Performance score seems to be very vague as around 75% of the population have the same performance score of 3 which seems to be a huge understatement as people in who need imporvement are still getting paid the same as those who exceed as per the data.

Finally, we can say that to predict a salary there are way more factors one has to look out for, other than the ones used here preferaably a higher volume of data too with more variance can be used to predict the model.