Федеральное государственное бюджетное образовательное учреждение

высшего профессионального образования

«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА и

ГОСУДАРСТВЕННОЙ СЛУЖБЫ

при Президенте Российской Федерации»

Институт бизнеса и делового администрирования

Программа магистратуры «Финансы и технологии»

**АНАЛИТИЧЕСКОЕ ЭССЕ**

по дисциплине “ **Интеллектуальный анализ данных: международный опыт и тренды**”

на тему:

**«Аналитический бизнес проект»**

**Автор работы**:

Студент 2 курса магистратуры

очной формы обучения

**Туа Эбенезер Квадво**

**Преподаватель:** Ирина Мухина, PhD

**МОСКВА**

**2019г.**

**Project Area**

When a US company wants to hire someone from outside of the United States for a technical position, they have to file an application to the United States government to get a green card or visa for the foreign applicant. These applications allow the US government to track who is entering and leaving the country for work-related reasons, and ensure that immigrants are neither being taken advantage of nor causing adverse effects for U.S. workers. To ensure equity for US and non-US workers, companies have to state how much they are planning on paying the employee every time they submit a visa or green card application. They also have to state the average amount an employee with similar skills and background typically gets paid for the same position, a figure called “the prevailing wage.” This publically available data provides a unique view into what types of salaries you might encounter for different data–related jobs in the US.

**Hypothesis:**

1. Are the average salaries of data-related jobs in Florida higher than that in Washington DC?
2. Is the average salary of a Data Scientist higher than Data Analyst in the US?

**Source of Data:**

The original data was compiled by the US Department of Labor’s Office of Foreign Labor Certification <http://www.foreignlaborcert.doleta.gov/performancedata.cfm>

<https://apps.bea.gov/iTable/index_regional.cfm>

Data Description(Metadata)

The data we used has a lot of observation of about 4349228 x 125 and we reduced it to 100 x 100

The variables are as follows:

|  |  |
| --- | --- |
| **FIELD NAME IN THE DATASET** | **DESCRIPTION** |
| Case Number | Unique identifier assigned to each application submitted for processing to the ETA National Processing Center. A = Atlanta NPC; C = Chicago NPC. |
| Case Status | Status associated with the last significant event or decision in the case application. Valid values include “Certified,” “Certified-Expired,” “Denied,” and “Withdrawn.” |
| Case Received Date | Date the application was received by the ETA National Processing Center. |
| Decision Date | Date on which the last significant event or decision was recorded by the ETA National Processing Center. |
| Employer Name | Name of employer requesting certification. |
| Prevailing Wage Submitted | Prevailing wage for the job being offered. |
|  |  |
| **FIELD NAME IN THE DATASET** | **DESCRIPTION** |
| Prevailing Wage Submitted Unit | Unit of Pay. Valid values include “Hourly (hr),” “Weekly (wk),” “Bi-Weekly (bi),” “Monthly (mth),” and “Yearly (yr).” |
| Paid Wage Submitted | Lower range of the wage offered to applicant |
| Paid Wage Submitted Unit | Unit of Pay. Valid values include “Hourly (hr),” “Weekly (wk),” “Bi-Weekly (bi),” “Monthly (mth),” and “Yearly (yr).” |
| Job Title | Payroll title of the job being offered, as entered on the application |
| Education Level Required | Minimum level of education required to adequately perform the duties of the job being offered |
| College Major Required | Major field of study required based on the education requirement. |
| Experience Required Y N | Identifies whether experience in the job offered is a requirement (Y/N). |
| Experience Required Num Months | The number of months experience that are required for the job (if applicable). |
| Country of Citizenship | Country of citizenship of the foreign worker being sponsored by the employer for employment in the United States. |
| Prevailing Wage Soc Code | Occupational code associated with the job being requested, as classified by the Standard Occupational Classification (SOC) System. |
| Prevailing Wage Soc Title | Name associated with the Prevailing Wage SOC Code. |
| Work City | City information of the foreign worker's intended area of employment. |
| Work State | State information of the foreign worker's intended area of employment. |
| Work Postal Code | Zip Code information of the foreign worker's intended area of employment |
| Full Time Position Y N | A binary variable indicating whether the position is full-time (Y) or part-time (N). |
| Visa Class | Indicates the type of application submitted for processing. Options include “H-1B,” “E-3 Australian,” “H-1B1 Chile,” “H-1B1 Singapore,” “greencard.” |
| Prevailing Wage Per Year | Prevailing wage per year, calculated for this Coursera course from the “Prevailing Wage Submitted” and “Prevailing Wage Submitted Unit” fields. |
| Paid Wage Per Year | Paid wage per year, calculated for this Coursera course from the “Paid Wage Submitted” and “Paid Wage Submitted Unit” fields. |
| Job Title Subgroup | Subgroup label made for this Coursera course based on the “Job Title” field. Options include assistant professor, attorney, business analyst, data analyst, data scientist, management consultant, software engineer, teacher. |

**Statistical hypothesis testing:**

The necessary columns of the data needed for the analysis basically showed a non-Gaussian distribution from the Anderson-Darling Normality Test and negative correlation from Pearson’s correlation test.

*Anderson-Darling Test*

Tests whether a data sample has a Gaussian distribution.

Assumptions

•Observations in each sample are independent and identically distributed (iid).

Interpretation

•H0: the sample has a Gaussian distribution.

•H1: the sample does not have a Gaussian distribution.

*Pearson’s Correlation Coefficient*

Tests whether two samples have a linear relationship.

Assumptions

Observations in each sample are independent and identically distributed (iid).

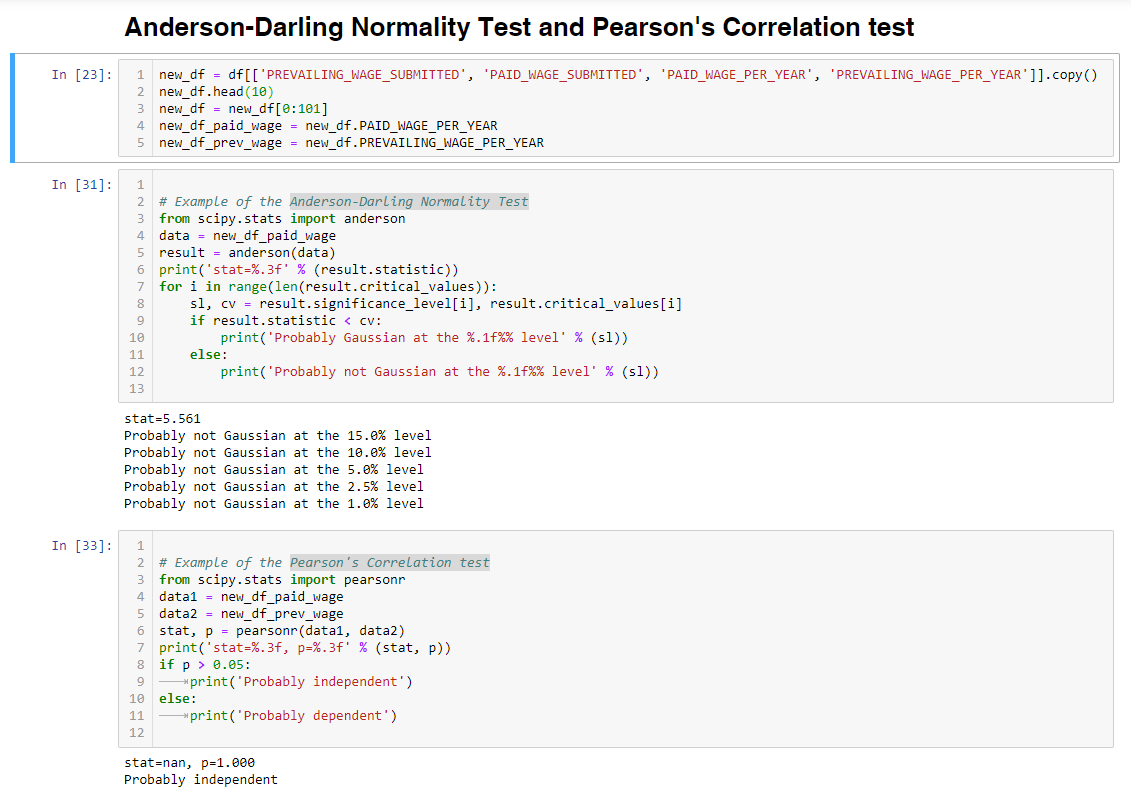
Observations in each sample are normally distributed.

Observations in each sample have the same variance.

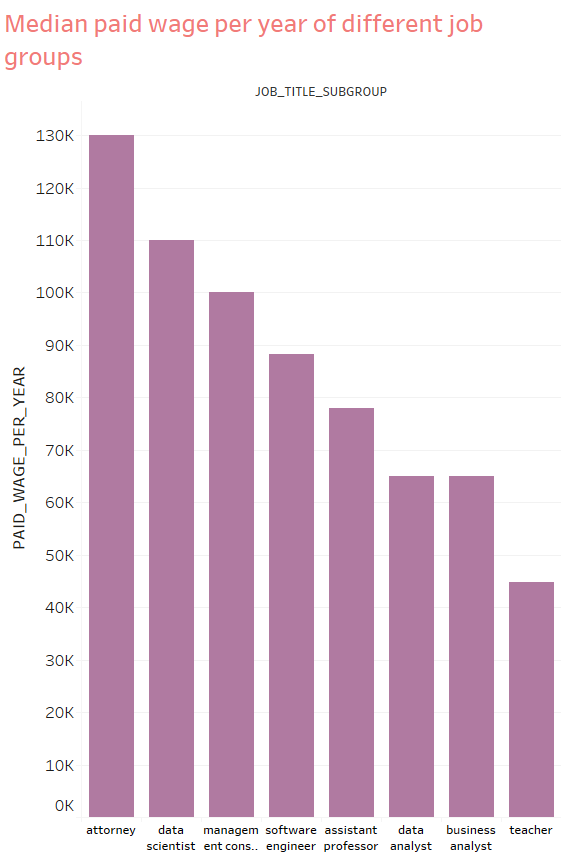
Interpretation

H0: the two samples are independent.

H1: there is a dependency between the samples.



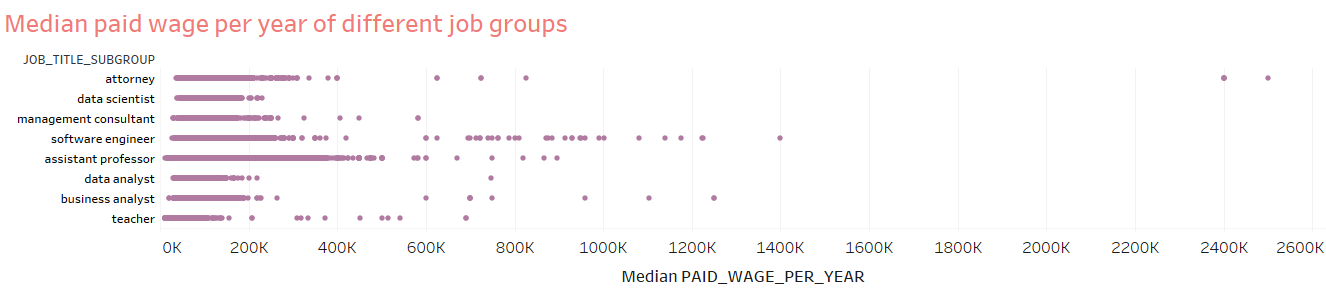
Finally, the Median Paid wage over the various Job groups has the distribution as below:



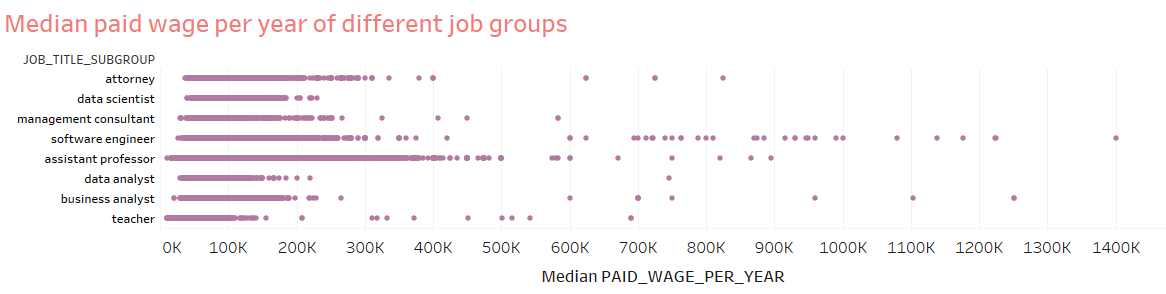
**Dealing with the Problems that the data had:**

* With the help of Big data tool from Google called Open Refine as was suggested to us during class provided a huge help to deal with the messy day. In there, wrongly spelt words were easily corrected and also pretty visible mistakes were detected and corrected.
* Dealing with outliers in the data. Our data upon in depth studies proofed to have some huge outliers amongst some of the job groups specifically the attorney subgroup and this was dealt with, with the help of Tableau. In there, we made a scatter plot of the data to help us see the outliers.

Before, it looked like this:

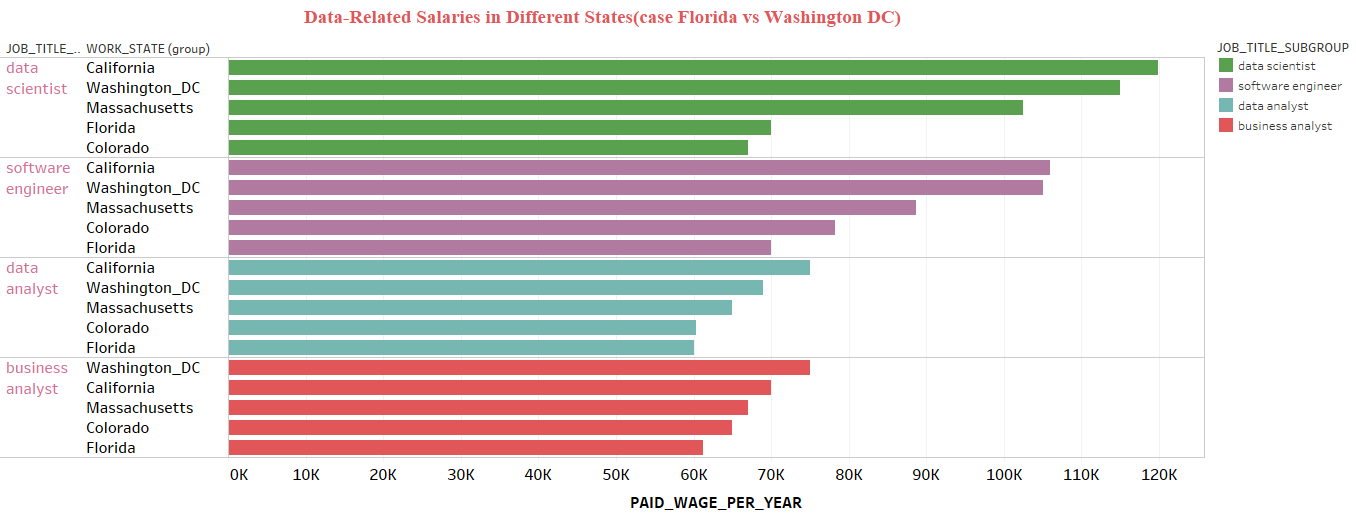


And then after removing them, we have:



**Analyzing Data-Related Salaries in Different States:**

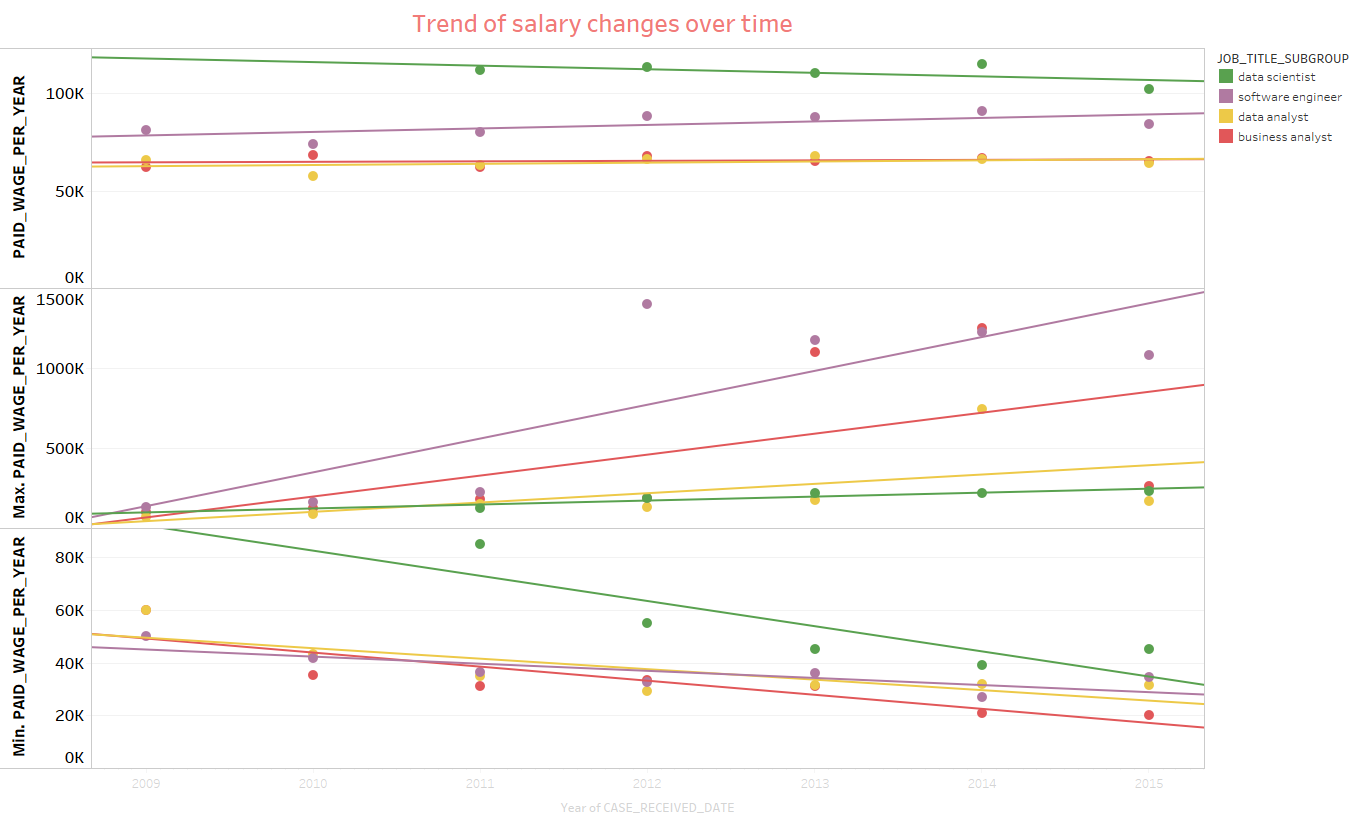
we did this by calculating the median of paid\_wage\_per\_year for each work\_state (group) broken down by job\_title\_subgroup. color shows details about job\_title\_subgroup. which is also represented in the legend found in the fig. below. the view is filtered on work\_state (group) and job\_title\_subgroup. the work\_state (group) filter keeps California, Colorado, Florida, Massachusetts and Washington dc. to make our observation easier we decided that the job\_title\_subgroup filter excludes assistant professor, attorney, management consultant and teacher.



We hence can see that California has almost the highest data-related job salaries followed by Washington DC amongst the state we choose for this analysis. It can be seen clearly that none of the data-related jobs offered in the State of Florida is higher than that of those in Washington DC and this can be easily deduced that perhaps Florida is not a high technological State as compared to Washington DC.

**Linear Regression into how the salaries may change over time:**

Since our analysis question allowed the initial use of 100 observations this will be extremely difficult to understand the trend line over the period of time. To any way try to see how the trend has been in the past, we obtained the same set of data from 2008 till date but took noticed till 2015 since it showed significant trend changes to try to understand how salaries in the data-related jobs have changed over time and we overserved that:



* For data for data scientist what we are seeing here indeed it does look like the maximum paid wage is going up over time. Even though this line looks like it's steeper than the one above it, it actually is not significant. It looks like for data scientist, the maximum paid wage is going up, but the median is not.
* For a software engineer, the maximum paid wage is going up over time, and the minimum paid wage is going down over time. Overall, again, the median paid wage is not changing.
* Data analyst can see that the minimum paid wage is going down overtime, whereas the maximum is staying basically the same. Again, the median is not changing.
* And for business analyst again, the minimum is going down, the maximum is stranding up, but it's not significant. And the median paid wage is not changing at all.

So according to this analysis it does look like the minimum salaries for Software Engineers, Data Analyst and Business Analyst are decreasing over time.

Even though the median paid wages are remaining exactly the same for all of these different categories.

The maximum paid wage for Data Scientists and Software Engineers on the other hand are increasing.

So overall that suggests that there might be increasing amount of opportunity to be awarded with very high salaries in data scientist jobs and software engineering jobs,

but overall the typical salaries for data related jobs seem to be staying quite constant. And for every category other than data scientists,

as time goes on, there might actually be increased risk that you will make less than you expected.

**Predicting the salaries of data-related jobs in the Future.**

In doing this part of the work, we used python to do Linear Regression of our data.

This was done by making our paid wage per year as the target and prevailing wage submitted, paid wage submitted, prevailing wage per year, visa class, job title subgroup as our features. And then data was split by 60% training set from the data and 20% as test set of the data. Then by the help of the model, we predicted the new paid wage per year.

The model gave the following results:

*R-square training*: 0.93

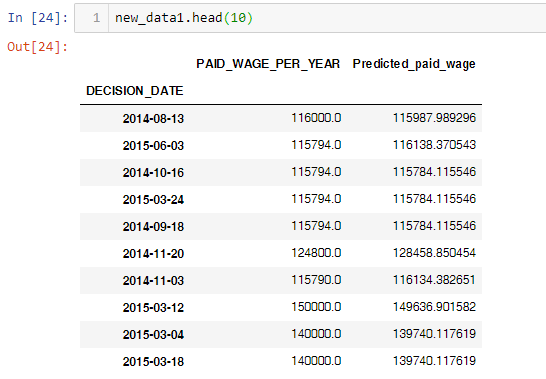
*R-square test*: 0.91

-------------------------------

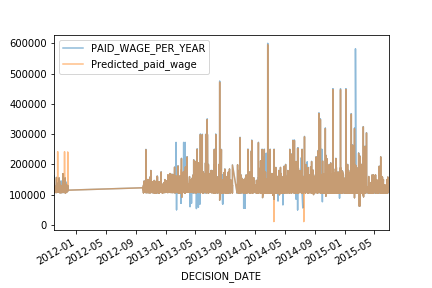
model.coef\_: [[-1.10320537 0.99080415 1.10971895 361.40250215 120.99298478]]

model. intercept\_: [-189.45641283]

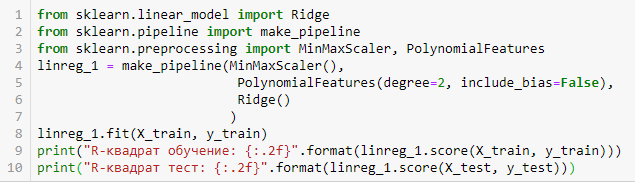
and a glance at the values it presented.



This again called for concern since the model seemed to be over fitted because of the high rate of accuracy score. We than

****

This again called for concern since the model seemed to be over fitted because of the high rate of accuracy score. We then used other instruments to help us very by using new model type after regularizing our data with the help of Ridge.

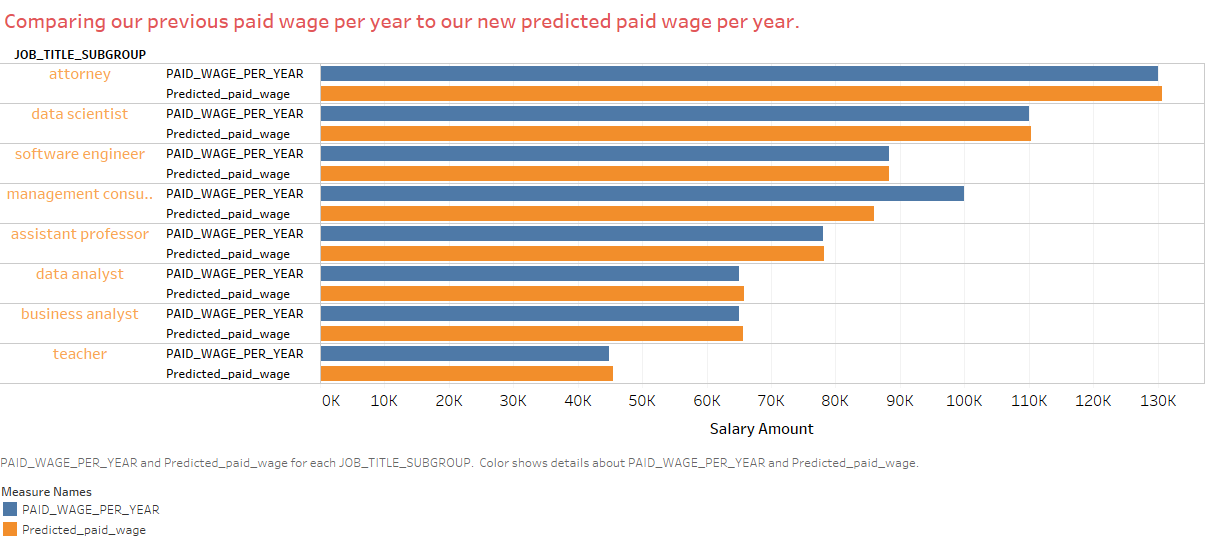


This anyway still gave us a high level of accuracy showing that the model is acceptable to use for the prediction.

*R-square training*: 0.93

*R-square test*: 0.91

Finally, with the help of Tableau, we visualize the difference between the previous paid wage per year and our new predicted paid wage per year. As seen in the figure below, we managed to predict very well amongst the various job subgroups.



**Conclusion of our analysis.**

We started our analysis with these hypothesis:

***H1***. Are the average salaries of data-related jobs in Florida higher than that in Washington DC?

***H2***. Is the average salary of a Data Scientist higher than Data Analyst in the US?

After our rigorous analysis we can comfortable affirm that;

***H1***. No, the average salaries of data-related jobs in Florida is not higher than that in Washington DC. It is the other way round.

***H2***. Yes, the average salary of a Data Scientist is higher than a Data Analyst in the US.