

salary?

CUNY MSDA DATA 606

Duubar Villalobos Jimenez

May 18, 2017

Part 1 - Introduction:

Nowadays with the increase in data collection and processing; companies, governments and agencies have a need to extract and produce educated decisions based on factual data. Extracting that kind of information and knowledge from large, heterogeneous, and noisy data sets requires not only powerful computing resources, but the programming abstractions to use them effectively.

In that context, data scientist need to have the skills in order to overcome the challenges that implies to work diverse structures in a given data science project.

By looking at this relationship (data processing / data scientist), we know that it creates a third component with a numerical variable named salary.

With that in mind, I will explore and try to answer a very important research question:

Are data science skills predictive of salary?

Hypothesis

From our exploration question, we can define our hypothesis as follows:

H_0 : Data Science skills are not predictive of salary; that is, the mean for all Skill Values are the same.

H_1 : Data Science skills are predictive of salary; that is, at least one mean for all Skill Values is different.

Part 2 - Data:

Data Source

The data that I will be working with, is collected by **Paysa** and is available online here: <http://paysa.com>



Figure 1:

For this project, the data was extracted by copying and pasting a job search of “*Data Science*” on March 16, 2017 into a text file, then cleaned and uploaded into a table in a remote MySQL server.

This data is collected by Paysa as part of the integrated job posting website and this data is submitted by employers daily.

Raw Data

The below table display all job listings compiled from Paysa.

```
## QApplication: invalid style override passed, ignoring it.  
## TypeError: Attempting to change the setter of an unconfigurable property.  
## TypeError: Attempting to change the setter of an unconfigurable property.
```

Cases

Each case represents a job posting in the United States. There are 390 observations in the given data set.

Explanatory variable

The explanatory variable is **Data Science skills** and is categorical.

Response variable

The response variable is **Base Salary** and is numerical.

Curated Data

From the above table I will focus on the **Base Salary** and combination of **Skills** as follows:

```
## QApplication: invalid style override passed, ignoring it.  
## TypeError: Attempting to change the setter of an unconfigurable property.  
## TypeError: Attempting to change the setter of an unconfigurable property.
```

Skill value per job listing

Since each case list multiple skills combined for a single base salary. For this study purposes, I will assign an “Skill Value” salary per skill listed on each listing; that is, by taking the base salary and dividing it by the number of skills listed for that study case.

For example: In the first case, there is a base salary of \$253000 with 6 skills listed (Distributed Systems, Big Data, Algorithms, Data Science, Strategy, Databases). By taking \$253000 and dividing it by 6, we obtain an average of \$42167. That is, each skill value will be taken as \$42167 in the first case study. Similar process will be applied for the rest of the cases.

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.  
## This warning is displayed once per session.
```

The below table shows the number of skills per job listing and also shows the “average” base salary for each skill in that listing.

```
## QApplication: invalid style override passed, ignoring it.  
## TypeError: Attempting to change the setter of an unconfigurable property.  
## TypeError: Attempting to change the setter of an unconfigurable property.
```

Part 3 - Exploratory data analysis:

From the above table, we have defined a series of `Skill Value` for each skill listed on each job posting.

From the raw data, we have a total of 2220 skills listed in the 390 job postings.

```
## QApplication: invalid style override passed, ignoring it.  
## TypeError: Attempting to change the setter of an unconfigurable property.  
## TypeError: Attempting to change the setter of an unconfigurable property.
```

Summary

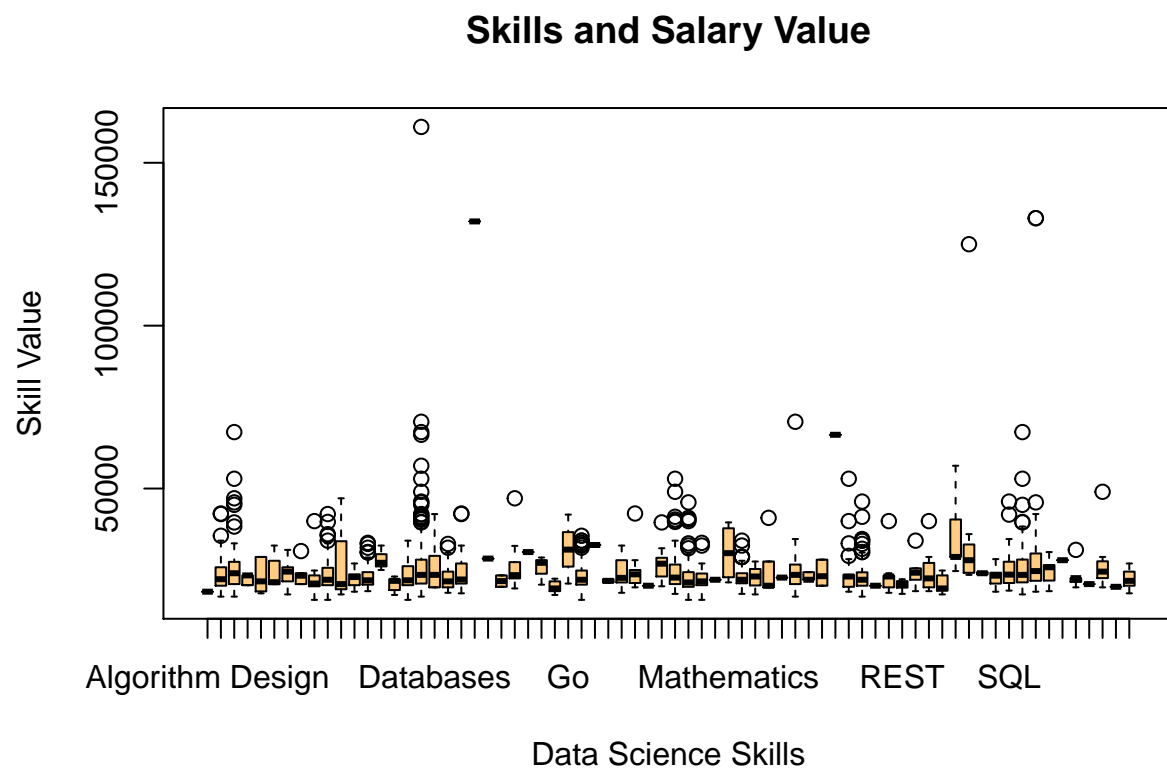
Below is a summary of the individual skills data.

```
##      Skills      Skill Value
## Length:2220    Min.   : 15833
## Class :character 1st Qu.: 20500
## Mode  :character Median : 22500
##                      Mean  : 24255
##                      3rd Qu.: 26617
##                      Max.   :161000
```

From the above summary table, we can quickly identify that the minimum skill value is set at \$15833 and the maximum is at \$161000 with a median skill value of \$22500 per skill.

Count, Mean and Standard deviation

```
## QApplication: invalid style override passed, ignoring it.
## TypeError: Attempting to change the setter of an unconfigurable property.
## TypeError: Attempting to change the setter of an unconfigurable property.
```

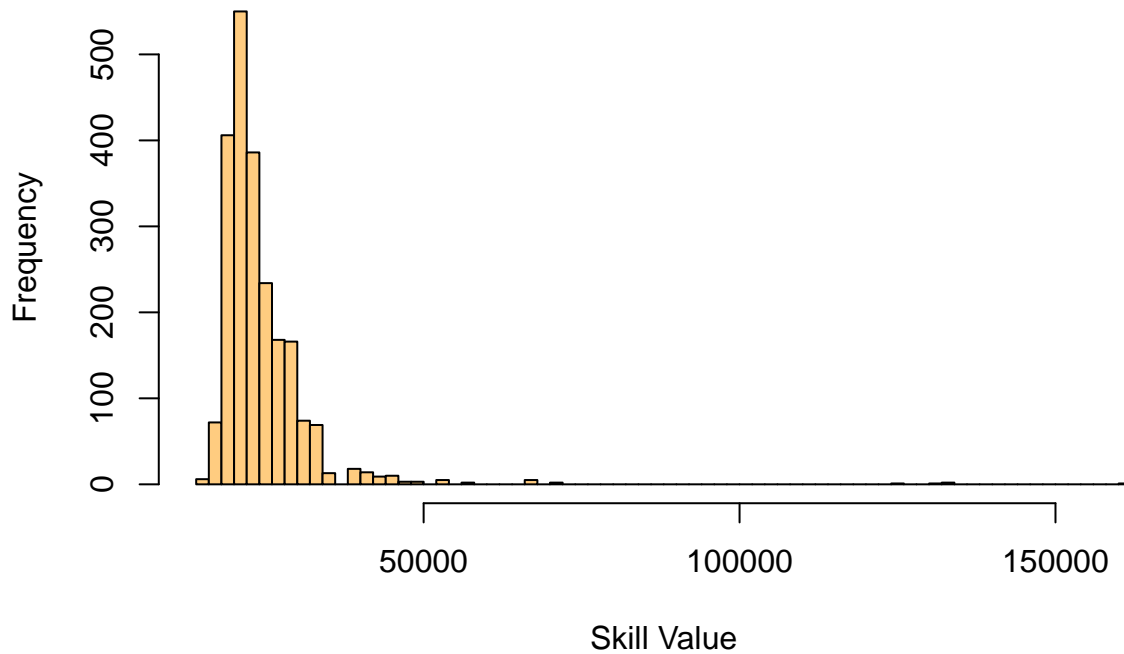


From the above plot, as an initial inspection of the data, it suggests that there are differences in between the medians but is not clear at this point.

Outliers

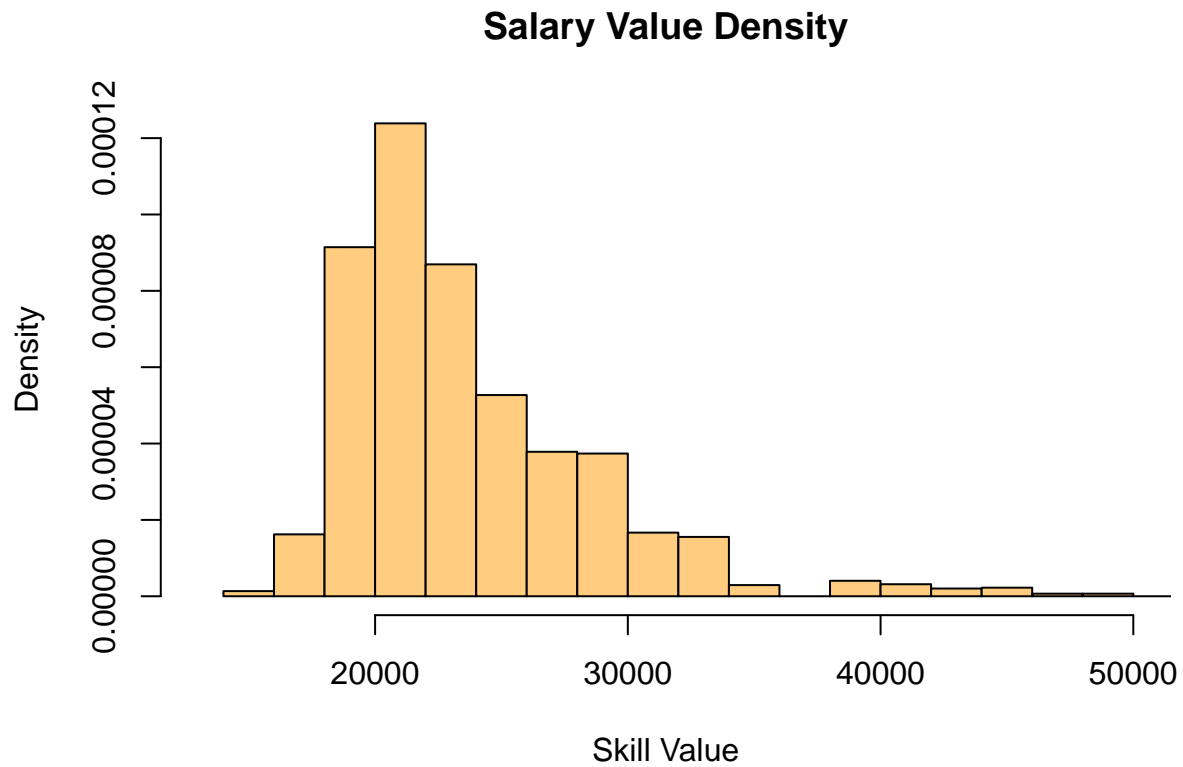
From the above set of box plots we can quickly identify that there are some outliers in the remaining data set, while the different medians seems to vary depending on the skill; this could be taking as an indication that the skills could be predictive of salary as established in our introduction.

Salary Value Frequency



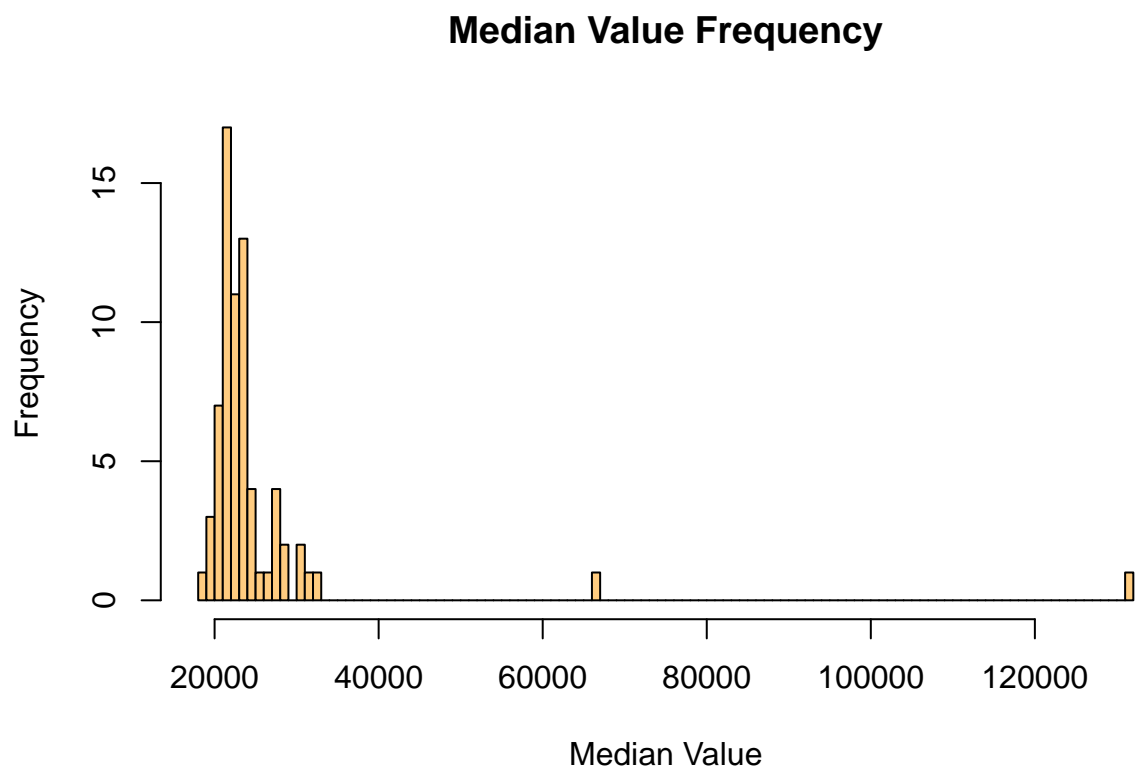
From the above histogram we can visualize some sort of normality and skewness to the right, also we can confirm the outliers as well.

For visualization purposes, I will include a new density histogram with a limited domain in the as follows:



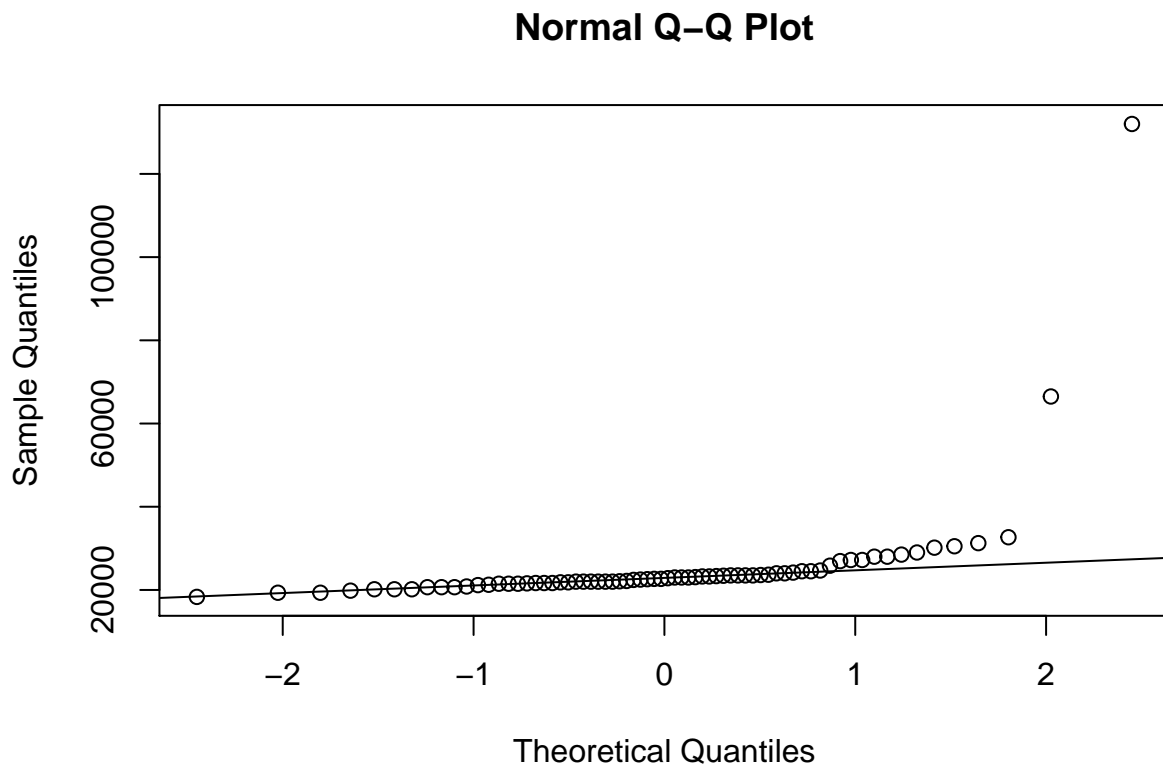
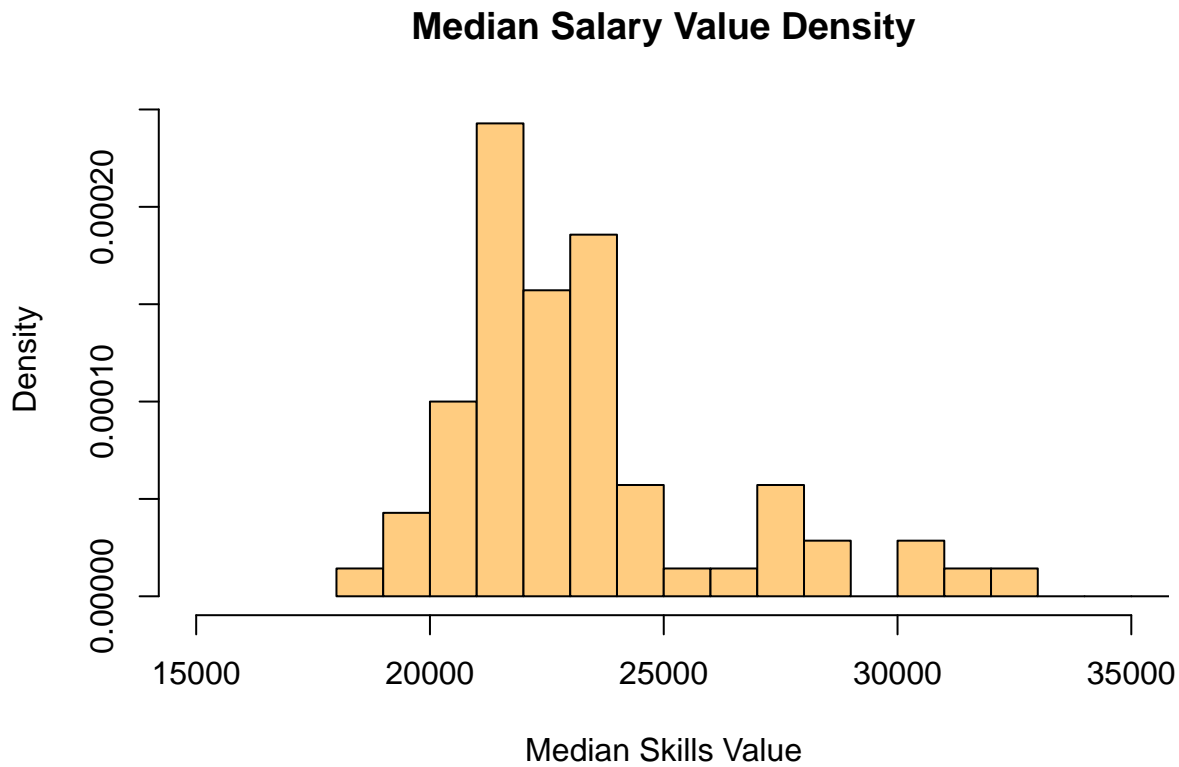
Medians

From the calculated medians, we can have the following histograms:



From the above histogram we can still visualize some sort of normality and skewness to the right, also we can confirm the presence of outliers, performing leverage.

For visualization purposes, I will include a new density histogram with a limited domain in the as follows:



Based on our Q-Q Plot, we can visualize how our medians data follow the qqline most of the trajectory then

due to leverage a couple of points fall away from it.

Part 4 - Inference:

Satisfying conditions for inference:

Conditions:

- The sample size is greater than 30.
- The data sets follow a uni modal normal distribution.
- The samples are random.

Hence, the conditions for inference seems to be satisfied.

ANOVA

Summary

```
##
## Call:
## lm(formula = `Skill Value` ~ Skills, data = my.skills.data_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17991  -3594  -1297   1854  134444
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   18333      7298    2.512  0.01208 *
## SkillsAlgorithms                5128      7318    0.701  0.48357
## SkillsAnalytics                 7499      7341    1.021  0.30715
## SkillsAndroid                  3800      7995    0.475  0.63463
## SkillsApache Spark             4750      7741    0.614  0.53952
## SkillsArchitecture             5695      7883    0.722  0.47013
## SkillsArchitectures            5387      7462    0.722  0.47048
## SkillsAutomation              5439      7883    0.690  0.49028
## SkillsAWS                     3873      7510    0.516  0.60613
## SkillsBig Data                 4928      7319    0.673  0.50085
## SkillsBusiness Intelligence   10056      8428    1.193  0.23291
## SkillsC                       3977      7802    0.510  0.61034
## SkillsC++                     4721      7350    0.642  0.52072
## SkillsCassandra               9900      8428    1.175  0.24022
## SkillsComputer Vision        2334      7655    0.305  0.76051
## SkillsData Mining            4729      7329    0.645  0.51881
## SkillsData Science           8223      7316    1.124  0.26113
## SkillsDatabases              7521      7741    0.972  0.33136
## SkillsDeep Learning          4380      7498    0.584  0.55920
## SkillsDistributed Systems     6723      7498    0.897  0.37007
## SkillsEMPTY                  113667    10322   11.013 < 2e-16 ***
## SkillsEngineering Management  10167    10322    0.985  0.32473
## SkillsEnterprise Software     3250      8939    0.364  0.71620
```

## SkillsETL	7509	7523	0.998	0.31836
## SkillsFirewalls	12167	10322	1.179	0.23861
## SkillsFunctional Programming	7178	8428	0.852	0.39446
## SkillsGame Development	1467	7995	0.183	0.85445
## SkillsGo	13028	8428	1.546	0.12228
## SkillsHadoop	4650	7318	0.635	0.52522
## SkillsHTTP	14334	10322	1.389	0.16506
## SkillsImage Processing	3334	8939	0.373	0.70920
## SkillsInformation Retrieval	5858	7498	0.781	0.43478
## SkillsJava	6439	7574	0.850	0.39534
## SkillsLAMP	1834	8939	0.205	0.83746
## SkillsLeadership	8739	7655	1.142	0.25374
## SkillsMachine Learning	5584	7312	0.764	0.44514
## SkillsManagement	5165	7358	0.702	0.48278
## SkillsMapReduce	4027	7415	0.543	0.58712
## SkillsMathematical Modeling	3667	10322	0.355	0.72242
## SkillsMathematics	11950	8160	1.464	0.14320
## SkillsMatlab	4940	7389	0.669	0.50383
## SkillsMySQL	4367	8428	0.518	0.60439
## SkillsNatural Language Processing	6378	7883	0.809	0.41855
## SkillsNetwork Architecture	4334	10322	0.420	0.67460
## SkillsOptimization	5755	7336	0.785	0.43278
## SkillsOS X	4600	7995	0.575	0.56510
## SkillsPHP	5473	7883	0.694	0.48761
## SkillsProduct Design	48167	10322	4.667	3.25e-06 ***
## SkillsProduct Management	5758	7408	0.777	0.43712
## SkillsPython	4771	7328	0.651	0.51510
## SkillsRecommender Systems	1834	8939	0.205	0.83746
## SkillsRelational Databases	4223	7538	0.560	0.57541
## SkillsREST	2122	7623	0.278	0.78079
## SkillsRuby	5462	7574	0.721	0.47091
## SkillsScala	5207	7394	0.704	0.48140
## SkillsScalability	2205	7574	0.291	0.77094
## SkillsScripting	17700	7995	2.214	0.02694 *
## SkillsSearch	23158	7802	2.968	0.00303 **
## SkillsSignal Processing	5667	8939	0.634	0.52616
## SkillsSoftware Design	4400	7655	0.575	0.56545
## SkillsSQL	6801	7363	0.924	0.35581
## SkillsStatistics	8266	7367	1.122	0.26195
## SkillsStrategy	16047	7432	2.159	0.03096 *
## SkillsTechnical Leadership	6123	7488	0.818	0.41362
## SkillsTest Driven Development	9667	10322	0.937	0.34908
## SkillsTime Series Analysis	4667	7802	0.598	0.54979
## SkillsTomcat	2334	8939	0.261	0.79403
## SkillsUser Experience	7855	7555	1.040	0.29856
## SkillsWeb Services	1500	10322	0.145	0.88447
## SkillsWindows	3941	7555	0.522	0.60199
## ---				
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
##				
## Residual standard error: 7298 on 2150 degrees of freedom				
## Multiple R-squared: 0.1714, Adjusted R-squared: 0.1448				
## F-statistic: 6.444 on 69 and 2150 DF, p-value: < 2.2e-16				

From the above results, the model output indicates some evidence of a difference in the average value for the skills.

Results

```
## Analysis of Variance Table
##
## Response: Skill Value
##           Df      Sum Sq   Mean Sq F value    Pr(>F)
## Skills      69 2.3685e+10 343255703   6.444 < 2.2e-16 ***
## Residuals 2150 1.1453e+11  53267887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the above table confirms that there are differences between the skills which were highlighted in the model summary.

Part 5 - Conclusion:

From our initial question: Are data science skills predictive of salary? we can conclude as follows:

By observing the above plots, linear modeling, and statistical analysis; we can observe how data science skills and income did appear to be correlated.

The validity of the data was indicated by summary statistics in which our hypothesis H_0 gets discarded and our alternative hypothesis H_1 is accepted. The above conclusion is statistically accepted since our analysis of variance returned an extremely low p-value (2.2e-16) which is less than 0.05. This can be enforced by comparing our results with the normality and qqplots for the medians as well.

References:

- OpenIntro Statistics, Third Edition. Diez, D. et all. 2015

Links

In order to open, right click and select **“Open Link in New Tab”**.

[dvillalobos.github.io](https://github.com/dvillalobos)

[GitHub](#) | [Linkedin](#)