

Data Project Proposal

CUNY MSDA DATA 606

Duubar Villalobos Jimenez

March 19, 2017

Final Project Proposal

Data Preparation

Let's take a look at our data; thus, to gain some initial insights.

```
# Reading our normalized data by employing RMySQL() in R
mydbconnection <- dbConnect(MySQL(),
                             user = myLocalUser,
                             password = myLocalPassword,
                             host = myLocalHost,
                             dbname = myLocalMySQLSchema)

# Check to see if our table exists? and read our dataset.
myLocalTableName <- tolower(myLocalTableName)
if (dbExistsTable(mydbconnection, name = myLocalTableName) == TRUE){
  my.data <- dbReadTable(mydbconnection, name = myLocalTableName)
} else {
  print("Error, the table does not exist")
}

# Closing connection with local Schema
dbDisconnect(mydbconnection)

#To close all open connections
lapply( dbListConnections( dbDriver( drv = "MySQL")), dbDisconnect)
```

Research question

Are Data Science skills predictive of salary?

Cases

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

Data collection

Data is collected by Paysa as part of the Integrated job posting website. Data is submitted by employers daily.

Type of study

This is an observational study.

Data Source

Data is collected by Paysa and is available online here: <http://paysa.com> For this project, data was extracted by copying and pasting a job search of “Data Science” on March 16, 2017 into a text file uploaded into a table in a remote MySQL server.

Response

The response variable is salary and is numerical.

Explanatory

The explanatory variable is Data Science skills and is categorical.

Relevant summary statistics

| Skills | Count | Percentage | Rank |
|-----------------------|-------|------------|------|
| Machine Learning | 241 | 10.87 % | 1 |
| Data Science | 201 | 9.07 % | 2 |
| Algorithms | 169 | 7.62 % | 3 |
| Hadoop | 153 | 6.9 % | 4 |
| Big Data | 152 | 6.86 % | 5 |
| Python | 116 | 5.23 % | 6 |
| Analytics | 85 | 3.83 % | 7 |
| Data Mining | 77 | 3.47 % | 8 |
| Optimization | 74 | 3.34 % | 9 |
| C++ | 69 | 3.11 % | 10 |
| SQL | 56 | 2.53 % | 11 |
| Management | 55 | 2.48 % | 12 |
| Statistics | 53 | 2.39 % | 13 |
| Data Mining | 42 | 1.89 % | 14 |
| Matlab | 40 | 1.8 % | 15 |
| Scala | 38 | 1.71 % | 16 |
| MapReduce | 31 | 1.4 % | 17 |
| Product Management | 30 | 1.35 % | 18 |
| Hadoop | 29 | 1.31 % | 19 |
| Strategy | 25 | 1.13 % | 20 |
| Big Data | 25 | 1.13 % | 20 |
| Optimization | 23 | 1.04 % | 22 |
| Architectures | 20 | 0.9 % | 23 |
| Machine Learning | 19 | 0.86 % | 24 |
| Deep Learning | 18 | 0.81 % | 26 |
| Distributed Systems | 18 | 0.81 % | 26 |
| AWS | 17 | 0.77 % | 27 |
| Information Retrieval | 16 | 0.72 % | 28 |
| ETL | 14 | 0.63 % | 30 |

| Skills | Count | Percentage | Rank |
|-----------------------------|-------|------------|------|
| User Experience | 14 | 0.63 % | 30 |
| Windows | 14 | 0.63 % | 30 |
| Algorithms | 14 | 0.63 % | 30 |
| Java | 13 | 0.59 % | 35 |
| Relational Databases | 13 | 0.59 % | 35 |
| Ruby | 13 | 0.59 % | 35 |
| Technical Leadership | 13 | 0.59 % | 35 |
| Scalability | 13 | 0.59 % | 35 |
| REST | 11 | 0.5 % | 38 |
| Computer Vision | 10 | 0.45 % | 40 |
| Leadership | 10 | 0.45 % | 40 |
| Apache Spark | 8 | 0.36 % | 42 |
| Databases | 8 | 0.36 % | 42 |
| Software Design | 8 | 0.36 % | 42 |
| C | 7 | 0.32 % | 45 |
| Time Series Analysis | 7 | 0.32 % | 45 |
| Python | 7 | 0.32 % | 45 |
| Architecture | 6 | 0.27 % | 50 |
| Natural Language Processing | 6 | 0.27 % | 50 |
| Search | 6 | 0.27 % | 50 |
| Data Science | 6 | 0.27 % | 50 |
| Management | 6 | 0.27 % | 50 |
| Technical Leadership | 6 | 0.27 % | 50 |
| Automation | 5 | 0.23 % | 54 |
| OS X | 5 | 0.23 % | 54 |
| Mathematics | 4 | 0.18 % | 56 |
| PHP | 4 | 0.18 % | 56 |
| Scripting | 4 | 0.18 % | 56 |
| Game Development | 4 | 0.18 % | 56 |
| Android | 3 | 0.14 % | 62 |
| Business Intelligence | 3 | 0.14 % | 62 |
| Cassandra | 3 | 0.14 % | 62 |
| Functional Programming | 3 | 0.14 % | 62 |
| Go | 3 | 0.14 % | 62 |
| MySQL | 3 | 0.14 % | 62 |
| Product Management | 3 | 0.14 % | 62 |
| Enterprise Software | 2 | 0.09 % | 73 |
| Image Processing | 2 | 0.09 % | 73 |
| LAMP | 2 | 0.09 % | 73 |
| Recommender Systems | 2 | 0.09 % | 73 |
| Signal Processing | 2 | 0.09 % | 73 |
| Tomcat | 2 | 0.09 % | 73 |
| Android | 2 | 0.09 % | 73 |
| Architectures | 2 | 0.09 % | 73 |
| C++ | 2 | 0.09 % | 73 |
| ETL | 2 | 0.09 % | 73 |
| Information Retrieval | 2 | 0.09 % | 73 |
| PHP | 2 | 0.09 % | 73 |
| Relational Databases | 2 | 0.09 % | 73 |
| Software Design | 2 | 0.09 % | 73 |
| Strategy | 2 | 0.09 % | 73 |
| Algorithm Design | 1 | 0.05 % | 88 |

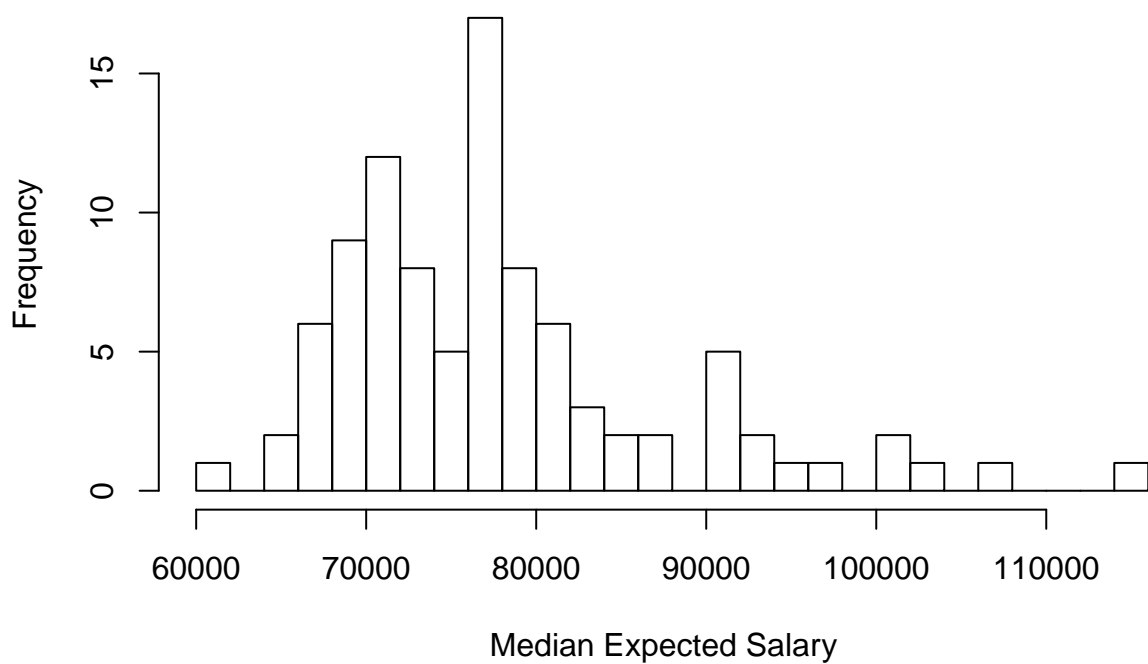
| Skills | Count | Percentage | Rank |
|-----------------------------|-------|------------|------|
| Engineering Management | 1 | 0.05 % | 88 |
| Firewalls | 1 | 0.05 % | 88 |
| Game Development | 1 | 0.05 % | 88 |
| HTTP | 1 | 0.05 % | 88 |
| Mathematical Modeling | 1 | 0.05 % | 88 |
| Network Architecture | 1 | 0.05 % | 88 |
| Test Driven Development | 1 | 0.05 % | 88 |
| Web Services | 1 | 0.05 % | 88 |
| Automation | 1 | 0.05 % | 88 |
| Data Science Scripting | 1 | 0.05 % | 88 |
| EMPTY | 1 | 0.05 % | 88 |
| Optimization Data Science | 1 | 0.05 % | 88 |
| Product Design Data Science | 1 | 0.05 % | 88 |
| Search | 1 | 0.05 % | 88 |

Top 10 paid Skills

| Skills | Type | Average | Median | Max | Min |
|---------------------------|-----------------|----------|--------|--------|--------|
| Automation | Expected Salary | 305000.0 | 305000 | 305000 | 305000 |
| Go | Expected Salary | 277333.3 | 277000 | 278000 | 277000 |
| Firewalls | Expected Salary | 255000.0 | 255000 | 255000 | 255000 |
| Technical Leadership | Expected Salary | 250666.7 | 254000 | 278000 | 231000 |
| Mathematics | Expected Salary | 226750.0 | 245500 | 287000 | 129000 |
| Engineering Management | Expected Salary | 231000.0 | 231000 | 231000 | 231000 |
| HTTP | Expected Salary | 229000.0 | 229000 | 229000 | 229000 |
| ETL | Expected Salary | 226000.0 | 226000 | 226000 | 226000 |
| Optimization Data Science | Expected Salary | 226000.0 | 226000 | 226000 | 226000 |
| ETL | Expected Salary | 216071.4 | 225500 | 255000 | 151000 |

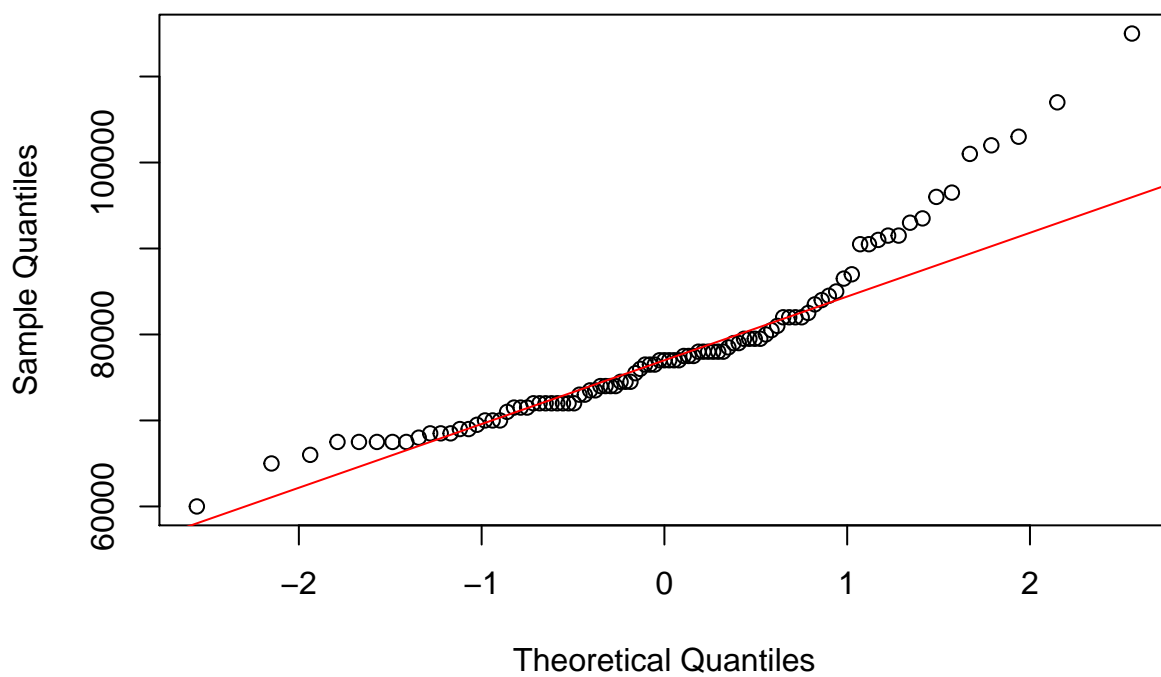
Salary Frequency

Median Salary Frequency



```
qqnorm(ind_salary_skills$Median)
qqline(ind_salary_skills$Median, col = 2)
```

Normal Q-Q Plot



Salary Type

| Type | Average | Median | Max | Min |
|-----------------|-----------|--------|--------|-------|
| Expected Salary | 176442.04 | 167000 | 338000 | 95000 |
| Base Salary | 138188.09 | 132000 | 265000 | 95000 |
| Annual Salary | 16169.15 | 18000 | 86000 | 0 |
| Signing Salary | 19047.36 | 17000 | 43000 | 0 |

Links

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

dvallalobos.github.io

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