# How to predict a salary for any US city

Baurjan Safi, DSI-5

General Assembly DC

# Plan of Presentation

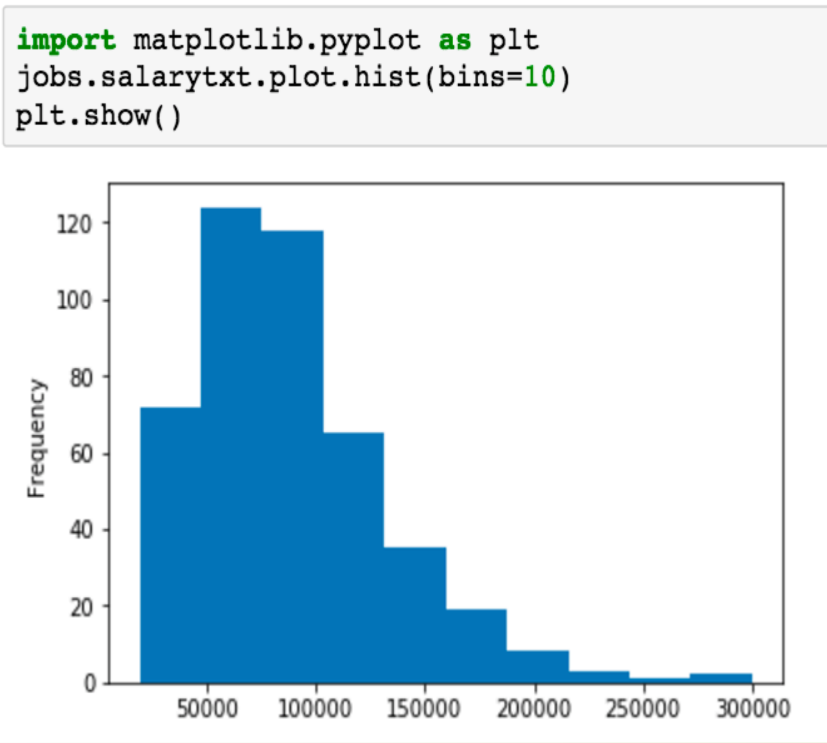
- Data Source and Data Set
- Other Data
- Identification of Key Factors
- Visuals and Rationale
- Model and Scores
- Words and Importance

# Data Source and Data Set

- Web-site: www.indeed.com
- Cities: 25
- 15 160 records, of them:
  - 8490 unique records
  - 6237 unique job titles
  - 3034 companies ( ~ 3 jobs per company)
  - 447 records with salaries (5.25%)
  - Salary range [$19,200 - $300,000*]

  * Simple Data Scientist at Intellipro Group, San Francisco, CA

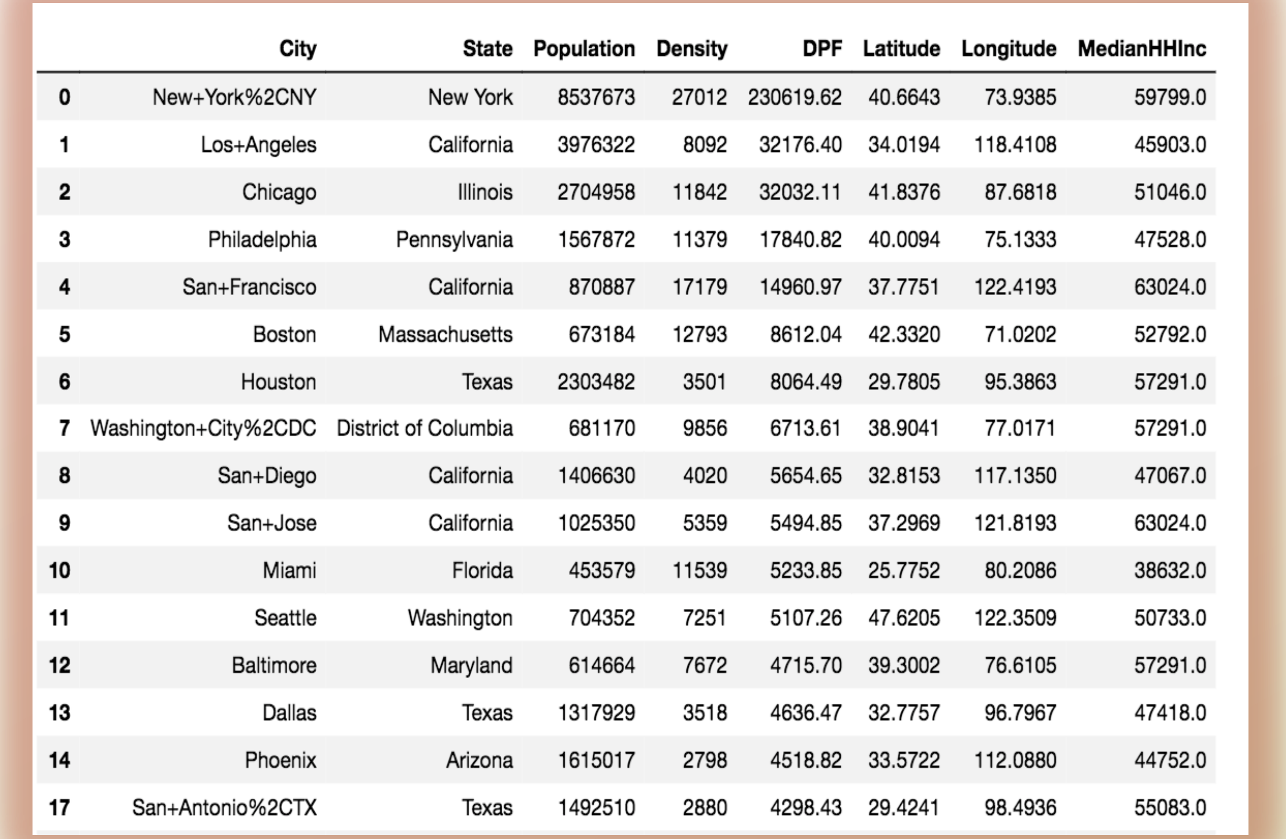
  Job Summary "…work with data engineers and other stakeholders in data products pipeline to enable automation of the data-driven products..."



```
import matplotlib.pyplot as plt
jobs.salarytxt.plot.hist(bins=10)
plt.show()
```

# Other Data

- City statistics:
  - 300 cities
  - Population
  - Density
  - Latitude
  - Longitude
  - Median Household Income



| | City | State | Population | Density | DPF | Latitude | Longitude | MedianHHInc |
|---|---|---|---|---|---|---|---|---|
| **0** | New+York%2CNY | New York | 8537673 | 27012 | 230619.62 | 40.6643 | 73.9385 | 59799.0 |
| **1** | Los+Angeles | California | 3976322 | 8092 | 32176.40 | 34.0194 | 118.4108 | 45903.0 |
| **2** | Chicago | Illinois | 2704958 | 11842 | 32032.11 | 41.8376 | 87.6818 | 51046.0 |
| **3** | Philadelphia | Pennsylvania | 1567872 | 11379 | 17840.82 | 40.0094 | 75.1333 | 47528.0 |
| **4** | San+Francisco | California | 870887 | 17179 | 14960.97 | 37.7751 | 122.4193 | 63024.0 |
| **5** | Boston | Massachusetts | 673184 | 12793 | 8612.04 | 42.3320 | 71.0202 | 52792.0 |
| **6** | Houston | Texas | 2303482 | 3501 | 8064.49 | 29.7805 | 95.3863 | 57291.0 |
| **7** | Washington+City%2CDC | District of Columbia | 681170 | 9856 | 6713.61 | 38.9041 | 77.0171 | 57291.0 |
| **8** | San+Diego | California | 1406630 | 4020 | 5654.65 | 32.8153 | 117.1350 | 47067.0 |
| **9** | San+Jose | California | 1025350 | 5359 | 5494.85 | 37.2969 | 121.8193 | 63024.0 |
| **10** | Miami | Florida | 453579 | 11539 | 5233.85 | 25.7752 | 80.2086 | 38632.0 |
| **11** | Seattle | Washington | 704352 | 7251 | 5107.26 | 47.6205 | 122.3509 | 50733.0 |
| **12** | Baltimore | Maryland | 614664 | 7672 | 4715.70 | 39.3002 | 76.6105 | 57291.0 |
| **13** | Dallas | Texas | 1317929 | 3518 | 4636.47 | 32.7757 | 96.7967 | 47418.0 |
| **14** | Phoenix | Arizona | 1615017 | 2798 | 4518.82 | 33.5722 | 112.0880 | 44752.0 |
| **17** | San+Antonio%2CTX | Texas | 1492510 | 2880 | 4298.43 | 29.4241 | 98.4936 | 55083.0 |

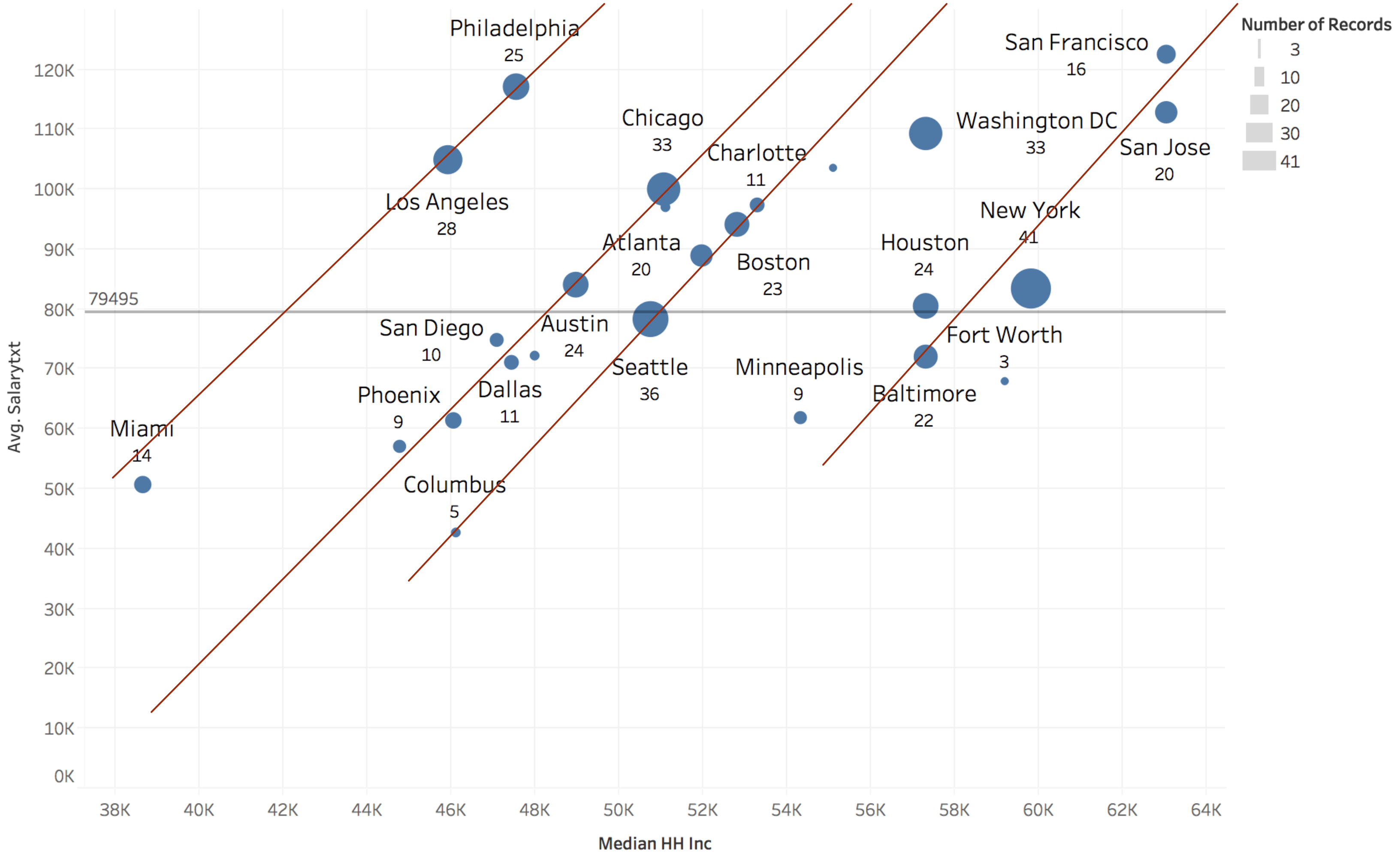
# Identification of Key Factors

- Population
- Location
- Density
- Density x Population Factor
- Median Household Income
- Average City Salary
- National Median Salary
- Words
- State

# Identification of Key Factors

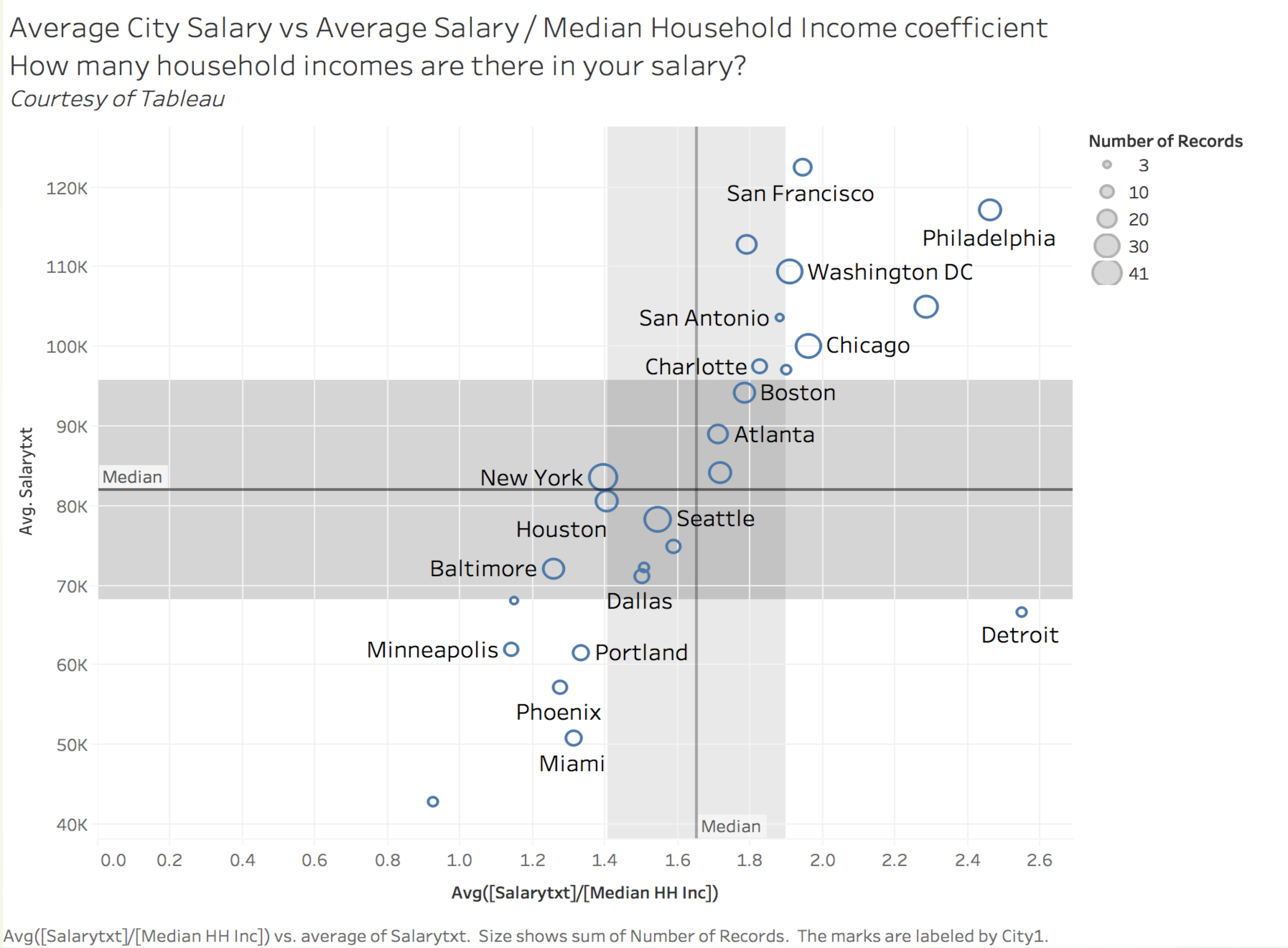
## Average Salary vs Median Household Income

*Courtesy of Tableau*



The trend of average of Salarytxt for Median HH Inc. Size shows sum of Number of Records. The marks are labeled by City1 and sum of Number of Records. The view is filtered on Exclusions (City,Median HH Inc), which keeps 25 members.

# Identification of Key Factors



Average City Salary vs Average Salary / Median Household Income coefficient
How many household incomes are there in your salary?
*Courtesy of Tableau*

Avg([Salarytxt]/[Median HH Inc]) vs. average of Salarytxt. Size shows sum of Number of Records. The marks are labeled by City1.

# Model and Scores

- Final Variables:
  - Coefficient of Salary vs Median Household Income
  - Population
  - Density
  - Words of Job Title (by presence of “president, senior, supervisor”
  - Words of Summary (unique 1800 words)
- Number of Categories:
  - 10
  - 4
  - 3
  - 2

# Importance of Variables

| Number of Categories | Salary to MHHI | Population | Density | MgrDummy |
|---|---|---|---|---|
| 10 | 0.808 | 0.072 | 0.094 | 0.026 |
| 4 | 0.847 | 0.057 | 0.063 | 0.033 |
| 3 | 0.838 | 0.06 | 0.068 | 0.034 |
| 2 | 0.923 | 0.02 | 0.028 | 0.029 |

# Models and Scores

| Number<br>Of categories | StratKFolds | Random Forest | RF with words | Bagging<br>Classifier |
|---|---|---|---|---|
| 10 | 0.654 | 0.638 | 0.44 | 0.641 |
| 4 | 0.794 | 0.801 | 0.723 | 0.796 |
| 3 | 0.863 | 0.863 | 0.81 | 0.857 |
| 2 | 0.926 | 0.931 | 0.926 | 0.931 |

# Words and Importance

| words | importance |
|---|---|
| data | 0.143575 |
| scientist | 0.066311 |
| research | 0.058424 |
| analysis | 0.057210 |
| analytics | 0.052634 |
| experience | 0.046657 |
| scientists | 0.042868 |
| looking | 0.040034 |
| analyze | 0.034816 |
| science | 0.030636 |

# Summary and Questions

- The most important variable – Salary / City Median Household Income
- The most accurate prediction model – Random Forest
- The most predictive variable split – two categories
- The best number of folds in Stratified K-fold – 10

- Questions?