# Salary and Wages Prediction

## How much should I pay as an employer for a specified occupation?

**Ella Huang** **28th-May-2020**

# Salary and Wages Prediction

## Example : Human Resources Manager

- Introduction
    - Problems and Data Description
- Methodology
- Results and Discussion
- Conclusion

# Introduction

## Problem and Data Description

A. The Problem:

- Jay is an employer who has the business size around  500 employees, his business focuses on semiconductor and electronic product manufacturing, his factories locate in the non-metro area, however, he would like to establish an office in New York City as the business grows. As an employer, Jay has two concerns about the recruitment of Human Resources Manager(HRM), 1. Labor cost 2.Competitive salaries and wages. Therefore , he'd like to know how much he should pay for this occupation , which is reasonable and maybe competitive in the labor market.

B. Data

- The data of Occupational Employment Statistics Survey from US Bureau of Labor Statistics will be used, and the website: www.bls.gov/oes for reference

# Methodology

- The tools include : Python, Pandas, Numpy, Matplotlib.
- The steps include:

1. Wrangling Data

2. Exploring Data

3.Model Development

4.Model Evaluation and Refinement

# Methdology

## 1.Wrangling Data

- Before wrangling the data, we should take a glance at the field description each and understand the features of the datasets.
- Check datasets shape( the number of rows and columns), information (i.e , the data types , if null values exist, and how many null values in each feature )
- Deal with null values, convert data types and pick up features related to the target

| | | | |
|---|---|---|---|
| 8 | Field | Field Description | NaN |
| 9 | area | U.S. (99), state FIPS code, Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA) code, or OES-specific nonmetropolitan area code | NaN |
| 10 | area_title | Area name | NaN |
| 11 | area_type | Area type: 1= U.S.; 2= State; 3= U.S. Territory; 4= Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA); 6= Nonmetropolitan Area | NaN |
| 12 | naics | North American Industry Classification System (NAICS) code for the given industry | NaN |
| 13 | naics_title | North American Industry Classification System (NAICS) title for the given industry | NaN |
| 14 | i_group | Industry level. Indicates cross-industry or NAICS sector, 3-digit, 4-digit, 5-digit, or 6-digit industry. For industries that OES no longer publishes at the 4-digit NAICS level, the “4-digit” designation indicates the most detailed industry breakdown available: either a standard NAICS 3-digit industry or an OES-specific combination of 4-digit industries. Industries that OES has aggregated to the 3-digit NAICS level (for example, NAICS 327000) will appear twice, once with the “3-digit” and once with the “4-digit” designation. | NaN |
| 15 | own_code | Ownership type: 1= Federal Government; 2= State Government; 3= Local Government; 123= Federal, State, and Local Government; 235=Private, State, and Local Government; 35 = Private and Local Government; 5= Private; 57=Private, Local Government Gambling Establishments (Sector 71), and Local Government Casino Hotels (Sector 72); 58= Private plus State and Local Government Hospitals; 59= Private and Postal Service; 1235= Federal, State, and Local Government and Private Sector | NaN |
| 16 | occ_code | The 6-digit Standard Occupational Classification (SOC) code or OES-specific code for the occupation | NaN |
| 17 | occ_title | SOC title or OES-specific title for the occupation | NaN |

# Methodology

## Exploring Data

- As you can see, we get a record of 395,647 rows and 31 columns in the datasets, and in view of the specified occupation is general and the industry requirement is not so significant , so we pick up the features including: area_title, occ_title,tot_emp,emp_prse, and datasets associated with average hourly wage and average annual wage.

[70]:

| | area_title | occ_title | tot_emp | emp_prse | h_mean | a_mean | mean_prse | h_pct10 | h_pct25 | h_median |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | U.S. | Management Occupations | 8054120 | 0.2 | 58.88 | 122480 | 0.1 | 24.03 | 34.35 | 50.80 |
| 1 | U.S. | Business and Financial Operations Occupations | 8183750 | 0.2 | 37.56 | 78130 | 0.2 | 18.76 | 25.06 | 33.57 |
| 2 | U.S. | Computer and Mathematical Occupations | 4552880 | 0.4 | 45.08 | 93760 | 0.5 | 21.79 | 30.22 | 42.47 |
| 3 | U.S. | Architecture and Engineering Occupations | 2592680 | 0.5 | 42.69 | 88800 | 0.3 | 21.77 | 29.28 | 39.15 |
| 4 | U.S. | Life, Physical, and Social Science Occupations | 1288920 | 0.7 | 37.28 | 77540 | 0.4 | 17.62 | 23.73 | 32.77 |

# Methodology

## Exploring Data

- After analysing the job of HRM, concluding the skills required and the job duties performed, we weight the job roles she/he would play, Administration(10%),Compensation and Benefits(20%),Human Resource General(50%),Training and Development (20%)

- Combined with the location information(New York-Newark-Jersey City), we filter 4 series,

  1. ASM(Administrative Service Managers)

  2. CBM(Compensation and Benefits Managers)

  3. HRM(Human Resources Managers)

  4. TDM(Training and Development Managers)

```
#TDM=AOE[AOE['occ_title']==n[5]].groupby('area_title').mean(
ASM
```

```
[171]:
tot_emp         25070.00
emp_prse            1.80
h_mean             67.69
a_mean         140800.00
mean_prse           0.90
h_pct10            37.09
h_pct25            48.35
h_median           61.67
h_pct75            79.07
h_pct90             0.00
a_pct10         77150.00
a_pct25        100570.00
a_median       128270.00
a_pct75        164460.00
a_pct90             0.00
Name: New York-Newark-Jersey City, NY-NJ-PA, dtype: float64
```

```
#TDM=AOE[AOE['occ_title']==n[5]].groupby('area_title').mean
CBM
```

```
[72]:
tot_emp          1650.00
emp_prse            4.10
h_mean             87.39
a_mean         181770.00
mean_prse           2.80
h_pct10            52.10
h_pct25            63.61
h_median           78.74
h_pct75             0.00
h_pct90             0.00
a_pct10        108370.00
a_pct25        132320.00
a_median       163780.00
a_pct75             0.00
a_pct90             0.00
Name: New York-Newark-Jersey City, NY-NJ-PA, dtype: float64
```

```
HRM
```

[73]:

```
tot_emp        12020.00
emp_prse           2.40
h_mean            81.77
a_mean        170070.00
mean_prse          3.60
h_pct10           42.31
h_pct25           54.26
h_median          73.87
h_pct75           98.99
h_pct90            0.00
a_pct10        88000.00
a_pct25       112870.00
a_median      153650.00
a_pct75       205900.00
a_pct90            0.00
Name: New York-Newark-Jersey City, NY-NJ-PA, dtype: float64
```

```
#TDM=AOE[AOE['occ_title']==n[5]].groupby('area_title').mean(
TDM
```

[74]:

```
tot_emp         3230.00
emp_prse           3.60
h_mean            80.70
a_mean        167850.00
mean_prse          1.20
h_pct10           46.48
h_pct25           60.46
h_median          75.98
h_pct75           94.95
h_pct90            0.00
a_pct10        96690.00
a_pct25       125750.00
a_median      158050.00
a_pct75       197500.00
a_pct90            0.00
Name: New York-Newark-Jersey City, NY-NJ-PA, dtype: float64
```

# Model Development

- Then,it's time to concatenate the four series (ASM,CBM,HRM,TDM),and give it a name Merged, as we mentioned the weights each role should take, so we create a list named Weighted to store the weights.
- Next, we create a column named Budgets to store the output of the model.
- Our model is the sum of the each feature multiplied by its weights .

[166]:

```
#Merged=pd.concat([ASM,CBM,HRM,TDM],axis=1)
#Merged.columns=n
#Merged['Budgets']=np.zeros(Merged.shape[0])
Weighted=[0.1,0.2,0.5,0.2]
a=[]
for x1,x2,x3,x4 in zip(Merged.iloc[:,0],Merged.iloc[:,1],Merged.iloc[:,2],Merged.iloc[:,3]):
    y=x1*Weighted[0]+x2*Weighted[1]+x3*Weighted[2]+x4*Weighted[3]
    a.append(y)
Merged['Budgets']=a
Merged
```

[166]:

| | Administrative Services and Facilities Managers | Compensation and Benefits Managers | Human Resources Managers | Training and Development Managers | Budgets |
|---|---|---|---|---|---|
| **tot_emp** | 25070.00 | 1650.00 | 12020.00 | 3230.00 | 9493.000 |
| **emp_prse** | 1.80 | 4.10 | 2.40 | 3.60 | 2.920 |
| **h_mean** | 67.69 | 87.39 | 81.77 | 80.70 | 81.272 |
| **a_mean** | 140800.00 | 181770.00 | 170070.00 | 167850.00 | 169039.000 |
| **mean_prse** | 0.90 | 2.80 | 3.60 | 1.20 | 2.690 |
| **h_pct10** | 37.09 | 52.10 | 42.31 | 46.48 | 44.580 |
| **h_pct25** | 48.35 | 63.61 | 54.26 | 60.46 | 56.779 |
| **h_median** | 61.67 | 78.74 | 73.87 | 75.98 | 74.046 |
| **h_pct75** | 79.07 | 0.00 | 98.99 | 94.95 | 76.392 |
| **h_pct90** | 0.00 | 0.00 | 0.00 | 0.00 | 0.000 |
| **a_pct10** | 77150.00 | 108370.00 | 88000.00 | 96690.00 | 92727.000 |
| **a_pct25** | 100570.00 | 132320.00 | 112870.00 | 125750.00 | 118106.000 |
| **a_median** | 128270.00 | 163780.00 | 153650.00 | 158050.00 | 154018.000 |
| **a_pct75** | 164460.00 | 0.00 | 205900.00 | 197500.00 | 158896.000 |
| **a_pct90** | 0.00 | 0.00 | 0.00 | 0.00 | 0.000 |

# Model Evaluation and Refinement

- Due to the model we build is for predicting the salary and wage for a specified occupation, and it's customised by cases , so we don't have uniformed data to evaluate and refine the model because the characteristics of each candidate is distinctive and unique. Therefore , this essential step is skipped for the model.

# Results and Discussion

- Now we can refer to the budgets for HRM, the mean of the hourly wage and the annual wage is $81.27 and $169,039 respectively, the median of the hourly wage and the annual wage is $74.046 and $154,018 respectively, also we can refer to the percentile of 10th, 25th, 50th, 75th when an employer considers how much reasonably should be paid. However, we can see the 90th is 0 ,which means that's no data available to predict it. Besides salaries and wages prediction, benefits we should also take into account when design the remuneration packages, benefits mainly covers: Insurance, holidays, vacations, retirement plan and savings, leave plans, and legally required benefits. As for insurance , especially health insurance which mainly include medical care, vision care, dental care, outpatient prescription drugs. In a word, we should keep an eye on the various factors that affect compensation and benefits, take care of the balance between the labor costs and remuneration packages.

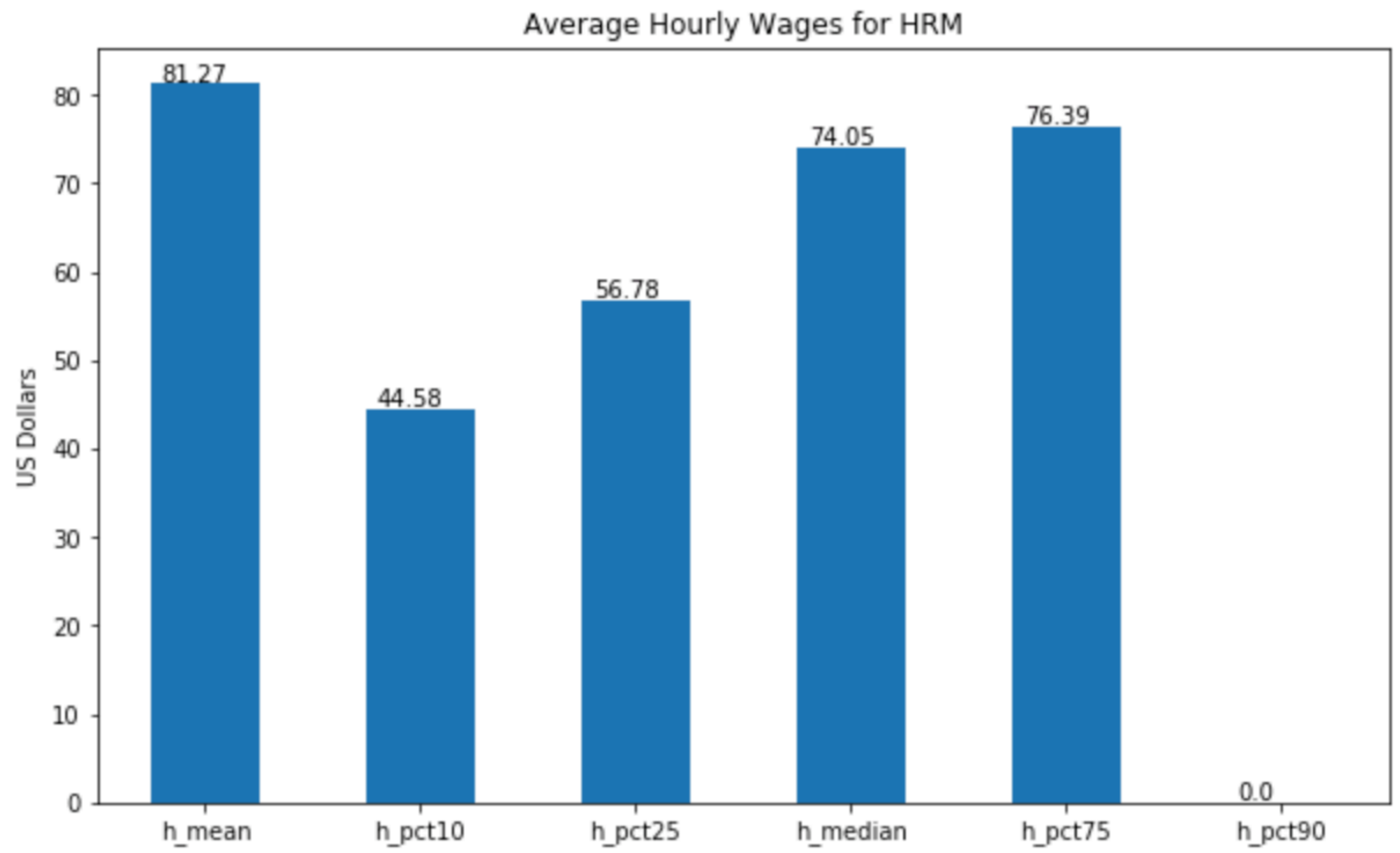
# Conclusion

- From the data analysis above and the diagram shows, we can conclude the points below for reference:

  1.The estimated average hourly wage for HRM in New York-Newark-Jersey City, USA is $81.27 and the median is $74.05

  2.The estimated average annually wage for HRM in New York-Newark-Jersey City, USA is $169,039 and the median is $154,018.

: <function matplotlib.pyplot.show(*args, **kw)>




.98]: <function matplotlib.pyplot.show(*args, **kw)>