Project 5_exploration_analysis

December 5, 2020

1 Data Visualization on Kaggle 2020

1.1 by Baiyan Ren

1.2 Preliminary Wrangling

This dataset is the annual survey of Kaggle on data science and machine learning in 2020. It collects the information of practitioners in a comprehensive way, from age, gender to prefered machine learning tools. I'll explore the dataset to understand the salary of data science and machine learning practitioners.

[2]:
```
# import all packages and set plots to be embedded inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

[3]:
```
survey_2020 = pd.read_csv('kaggle_survey_2020_responses.csv', low_memory=False,␣
↪skiprows=[1])
```

[16]:
```
survey_2020.head()
```

[16]:
```
   Time from Start to Finish (seconds)    Age Gender  \
0                                 1838  35-39    Man
1                               289287  30-34    Man
2                                  860  35-39    Man
3                                  507  30-34    Man
4                                   78  30-34    Man

                   Country          Education              Title  \
0                 Colombia    Doctoral degree            Student
1  United States of America    Master's degree      Data Engineer
2                Argentina  Bachelor's degree  Software Engineer
3  United States of America    Master's degree     Data Scientist
4                    Japan    Master's degree  Software Engineer
```

```
    Coding_exp Q7_Part_1 Q7_Part_2 Q7_Part_3  … Q35_B_Part_2 Q35_B_Part_3  \
0   5-10 years    Python         R       SQL  …          NaN          NaN
1   5-10 years    Python         R       SQL  …          NaN          NaN
2  10-20 years       NaN       NaN       NaN  …          NaN          NaN
3   5-10 years    Python       NaN       SQL  …          NaN          NaN
4    3-5 years    Python       NaN       NaN  …          NaN          NaN

  Q35_B_Part_4   Q35_B_Part_5 Q35_B_Part_6 Q35_B_Part_7 Q35_B_Part_8  \
0          NaN   TensorBoard          NaN          NaN          NaN
1          NaN            NaN          NaN          NaN          NaN
2          NaN            NaN          NaN          NaN          NaN
3          NaN            NaN          NaN          NaN          NaN
4          NaN            NaN          NaN          NaN          NaN

  Q35_B_Part_9 Q35_B_Part_10 Q35_B_OTHER
0          NaN           NaN         NaN
1          NaN           NaN         NaN
2          NaN          None         NaN
3          NaN           NaN         NaN
4          NaN           NaN         NaN

[5 rows x 355 columns]
```

[3]:
```
survey_2020.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20036 entries, 0 to 20035
Columns: 355 entries, Time from Start to Finish (seconds) to Q35_B_OTHER
dtypes: int64(1), object(354)
memory usage: 54.3+ MB
```

[4]:
```
survey_2020.isna().sum()
```

[4]:
```
Time from Start to Finish (seconds)        0
Q1                                         0
Q2                                         0
Q3                                         0
Q4                                       467
                                        …
Q35_B_Part_7                           19556
Q35_B_Part_8                           19190
Q35_B_Part_9                           19517
Q35_B_Part_10                          16954
Q35_B_OTHER                            19785
Length: 355, dtype: int64
```

1.2.1 What is the structure of your dataset?

It has 20036 rows and 355 colomns

1.2.2 What is/are the main feature(s) of interest in your dataset?

salary

1.2.3 What features in the dataset do you think will help support your investigation into your feature(s) of interest?

columns containing the information of education level and coding experience

1.3 Univariate Exploration

[4]:
```
columns = {'Q1': 'Age',
           'Q2': 'Gender',
           'Q3': 'Country',
           'Q4': 'Education',
           'Q5': 'Title',
           'Q6': 'Coding_exp',
           'Q8': 'Recommended_language',
           'Q15': 'ML_exp',
           'Q20': 'Company_size',
           'Q24': 'Salary'}
survey_2020.rename(columns=columns, inplace=True)
```

[5]:
```
Education = ['No formal education past high school',
             'Some college/university study without earning a bachelor's␣
 ↪degree',
             'Bachelor's degree',
             'Master's degree',
             'Doctoral degree',
             'Professional degree',
             'I prefer not to answer']
edu = pd.api.types.CategoricalDtype(categories=Education, ordered=True)
survey_2020['Education'] = survey_2020['Education'].astype(edu)
```

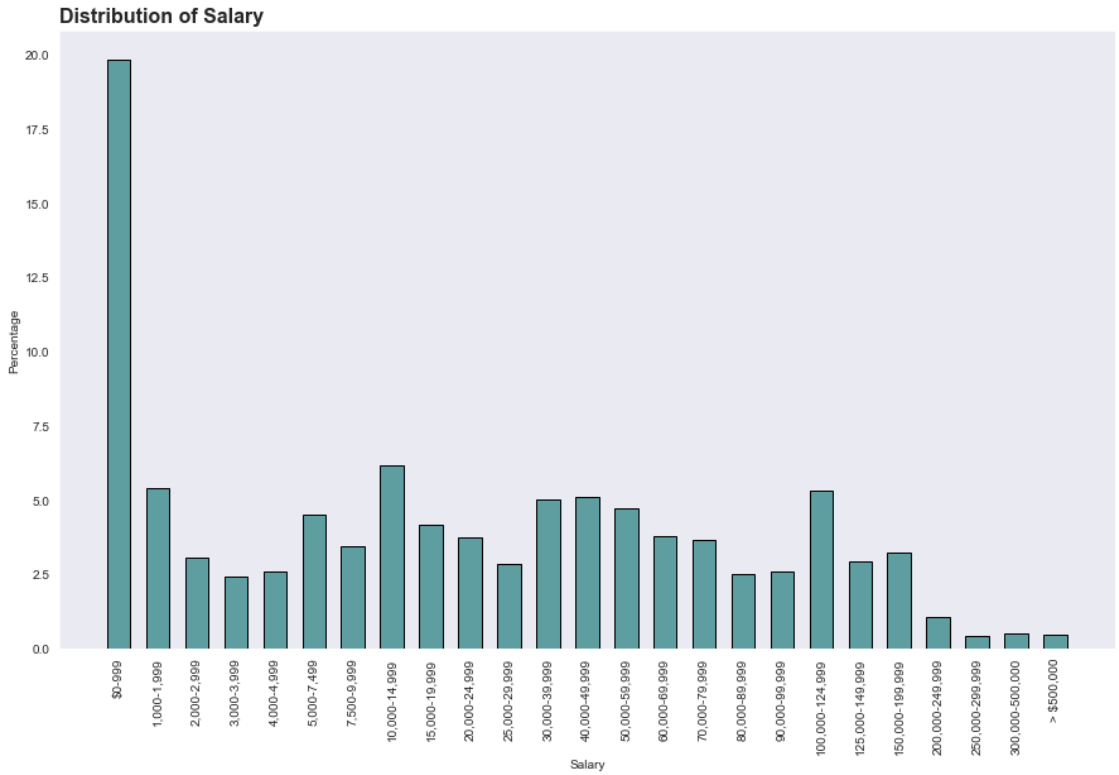
[6]:
```
survey_2020_doct = survey_2020.query('Education == "Doctoral degree"').copy()
survey_2020_other = survey_2020.query('Education != "Doctoral degree"').copy()
```

[7]:
```
edu_count = survey_2020.groupby('Education').size()
total = edu_count.sum()
edu_count = edu_count/total*100
edu_count
```

[7]:
```
Education
No formal education past high school                                  1.226430
Some college/university study without earning a bachelor's degree     5.580254
Bachelor's degree                                                    35.658439
Master's degree                                                      40.160458
Doctoral degree                                                      11.763504
Professional degree                                                   3.571976
I prefer not to answer                                                2.038939
dtype: float64
```

[8]:
```
salary = ['$0-999', '1,000-1,999', '2,000-2,999', '3,000-3,999', '4,000-4,999',␣
↪'5,000-7,499', '7,500-9,999',
          '10,000-14,999', '15,000-19,999', '20,000-24,999', '25,000-29,999',␣
↪'30,000-39,999', '40,000-49,999',
          '50,000-59,999', '60,000-69,999', '70,000-79,999', '80,000-89,999',␣
↪'90,000-99,999', '100,000-124,999',
          '125,000-149,999', '150,000-199,999', '200,000-249,999',␣
↪'250,000-299,999', '300,000-500,000', '> $500,000']
salary_total = survey_2020.groupby('Salary').size()[salary]
salary_prop_total = salary_total/salary_total.sum()*100
```

[9]:
```
sns.set_style('dark')
fig, ax = plt.subplots(figsize=[15, 9])
ax.bar(salary_total.index, salary_prop_total, color='cadetblue', edgecolor=(0,␣
↪0, 0), width=0.6, label='All')
plt.xticks(rotation=90)
plt.xlabel('Salary')
plt.ylabel('Percentage')
plt.title('Distribution of Salary', fontsize=16, fontweight='bold', loc='left');
```



Majority of the respondents have annual salary lower than $1000.

Next, I'll explore the distribution of salary and coding experience.

1.3.1 Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

Majority of the respondents have annual salary lower than \$1000, which is unusual in US. The reason might be that this is collected from worldwide, the salary is different in developed and developing country.

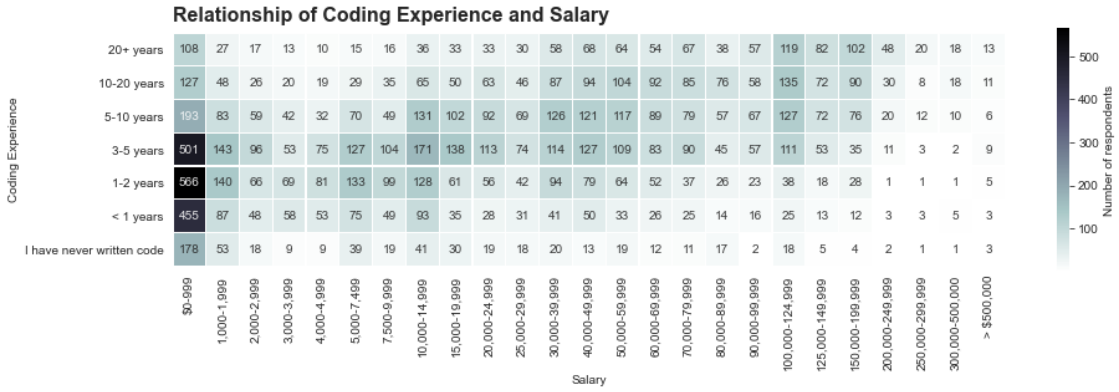
1.3.2 Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

No

1.4 Bivariate Exploration

[10]:
```
coding = ['I have never written code', '< 1 years', '1-2 years', '3-5 years',␣
 ↪'5-10 years', '10-20 years', '20+ years']
coding_salary = survey_2020.groupby(['Coding_exp', 'Salary']).size().
 ↪unstack()[salary].reindex(coding[::-1]).fillna(0).astype(int)
```

[11]:
```
fig, ax = plt.subplots(figsize=[15, 9])

sns.heatmap(data=coding_salary,
                    cmap='bone_r',
                    linewidths=0.2,
                    square=True,
                    annot=True,
                    fmt = 'd',
                    annot_kws={'alpha': 0.9},
                    cbar_kws={'shrink': 0.4, 'label': 'Number of respondents'})
plt.xlabel('Salary')

plt.ylabel('Coding Experience')

plt.title('Relationship of Coding Experience and Salary', fontsize=16,␣
 ↪fontweight='bold', loc='left', va='bottom');
```



There is a positive correlation between coding experience and salary. Next, I'll bring the third variable, education level.

1.4.1 Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Although most of the respondents have salary lower than \$1000, there is a positive correlation between salary and coding experience.

1.4.2 Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

No

1.5 Multivariate Exploration

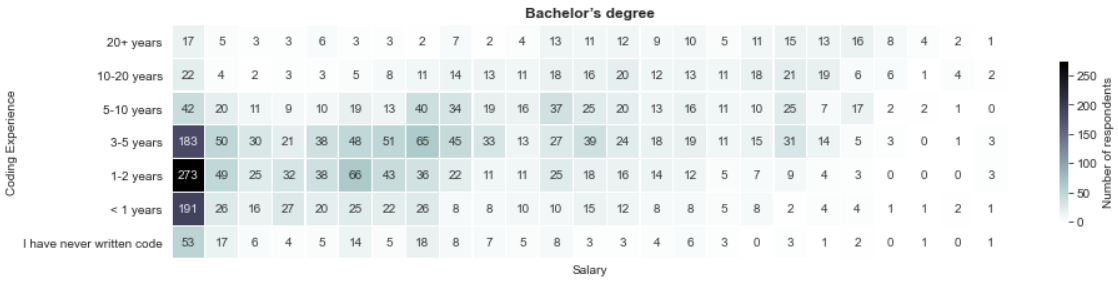
[26]:
```
survey_2020_bs = survey_2020.query('Education == "Bachelor's degree"').copy()
survey_2020_ms = survey_2020.query('Education == "Master's degree"').copy()
```

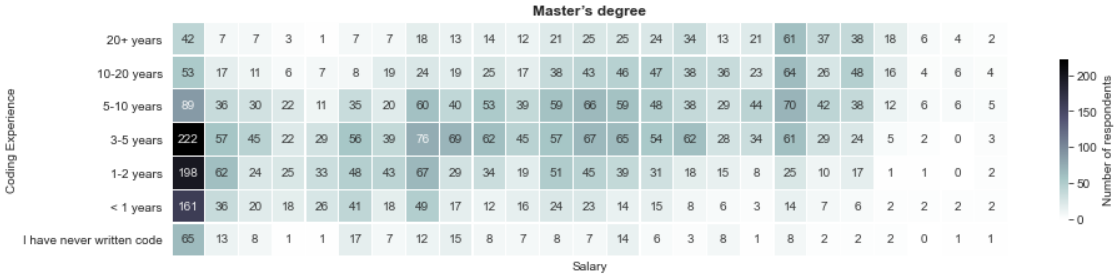
[36]:
```
coding_bs = survey_2020_bs.groupby(['Coding_exp', 'Salary']).size().
 ↪unstack()[salary].reindex(coding[::-1]).fillna(0).astype(int)
coding_ms = survey_2020_ms.groupby(['Coding_exp', 'Salary']).size().
 ↪unstack()[salary].reindex(coding[::-1]).fillna(0).astype(int)
coding_doct = survey_2020_doct.groupby(['Coding_exp', 'Salary']).size().
 ↪unstack()[salary].reindex(coding[::-1]).fillna(0).astype(int)
```

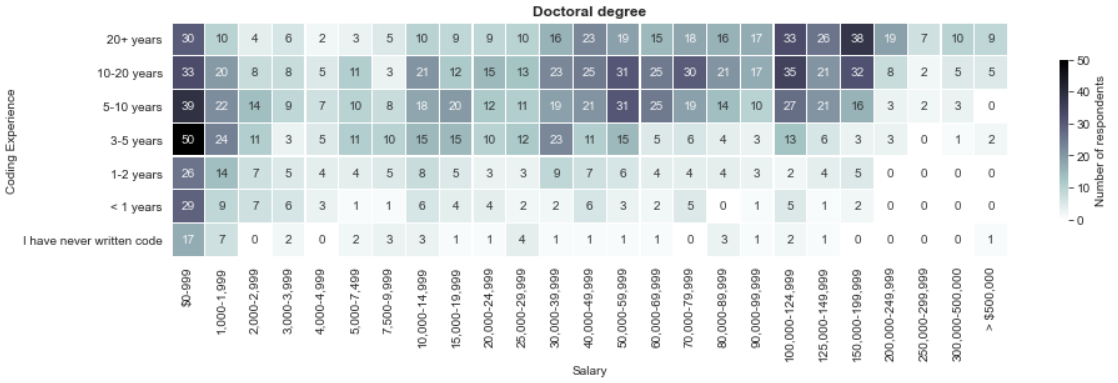
[57]:
```
fig, axes = plt.subplots(3, 1, figsize=[15, 20], sharex=True, sharey=True)

ax1 = sns.heatmap(data=coding_bs,
                  cmap='bone_r',
                  linewidths=0.2,
                  square=True,
                  annot=True,
                  fmt = 'd',
                  annot_kws={'alpha': 0.9},
                  cbar_kws={'shrink': 0.4, 'label': 'Number of respondents'},
                  ax=axes[0],
                  label='Bachelor's degree')
ax2 = sns.heatmap(data=coding_ms,
                  cmap='bone_r',
                  linewidths=0.2,
                  square=True,
                  annot=True,
                  fmt = 'd',
                  annot_kws={'alpha': 0.9},
                  cbar_kws={'shrink': 0.4, 'label': 'Number of respondents'},
                  ax=axes[1],
                  label='Master's degree')
ax3 = sns.heatmap(data=coding_doct,
                  cmap='bone_r',
                  linewidths=0.2,
                  square=True,
                  annot=True,
                  fmt = 'd',
                  annot_kws={'alpha': 0.9},
```

```
                  cbar_kws={'shrink': 0.4, 'label': 'Number of respondents'},
                  ax=axes[2],
                  label='Doctoral degree')
fontdict={'fontsize': 12,
          'fontweight': 'bold'}
ax1.set_title('Bachelor's degree', fontdict=fontdict)
ax2.set_title('Master's degree', fontdict=fontdict)
ax3.set_title('Doctoral degree', fontdict=fontdict)
for ax in [ax1, ax2, ax3]:
    ax.set_ylabel('Coding Experience')

fig.suptitle('Relationship of Coding Experience and Salary', x=0.3, y=0.9,␣
 ↪size=20, weight='bold');
```

Relationship of Coding Experience and Salary







1.5.1 Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

There is positive correlation between salary and coding experience in all three education levels. In addition, with the same coding experience, the salary of respondents with doctoral degree is higher than those with Master's degree and Bachelor's degree. ### Were there any interesting or surprising interactions between features?

The proportion of experienced respondents increases with education level.

[ ]: