Midterm Assignment

Note

You can find the `blizzard_salary.csv` data here.

In 2020, employees of Blizzard Entertainment circulated a spreadsheet to anonymously share salaries and recent pay increases amidst rising tension in the video game industry over wage disparities and executive compensation. (Source: Blizzard Workers Share Salaries in Revolt Over Pay)

The name of the data frame used for this analysis is `blizzard_salary` and the variables are:

- `percent_incr`: Raise given in July 2020, as percent increase with values ranging from 1 (1% increase to 21.5 (21.5% increase)
- `salary_type`: Type of salary, with levels `Hourly` and `Salaried`
- `annual_salary`: Annual salary, in USD, with values ranging from $50,939 to $216,856.
- `performance_rating`: Most recent review performance rating, with levels `Poor`, `Successful`, `High`, and `Top`. The `Poor` level is the lowest rating and the `Top` level is the highest rating.

The top ten rows and `.info` of `blizzard_salary` are shown below:

```
   percent_incr salary_type  annual_salary performance_rating
0           1.0        year            1.0               High
1           1.0        year            1.0         Successful
2           1.0        year            1.0               High
3           1.0      Hourly        33987.2         Successful
4           NaN      Hourly        34798.4               High
5           NaN      Hourly        35360.0                NaN
6           NaN      Hourly        37440.0                NaN
7           0.0      Hourly        37814.4                NaN
8           4.0      Hourly        41100.8                Top
```

```
9           1.2      Hourly       42328.0                NaN
<class 'pandas.core.frame.DataFrame'>
Index: 409 entries, 0 to 465
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   percent_incr        370 non-null    float64
 1   salary_type         409 non-null    object
 2   annual_salary       409 non-null    float64
 3   performance_rating  298 non-null    object
dtypes: float64(2), object(2)
memory usage: 16.0+ KB
None
```

Question 1

Which of the following is **correct**? Choose all that apply.

a. The `blizzard_salary` dataset has 399 rows.

b. The `blizzard_salary` dataset has 4 columns.

c. Each row represents a Blizzard Entertainment worker who filled out the spreadsheet.

d. The `percent_incr` variable is numerical and discrete.

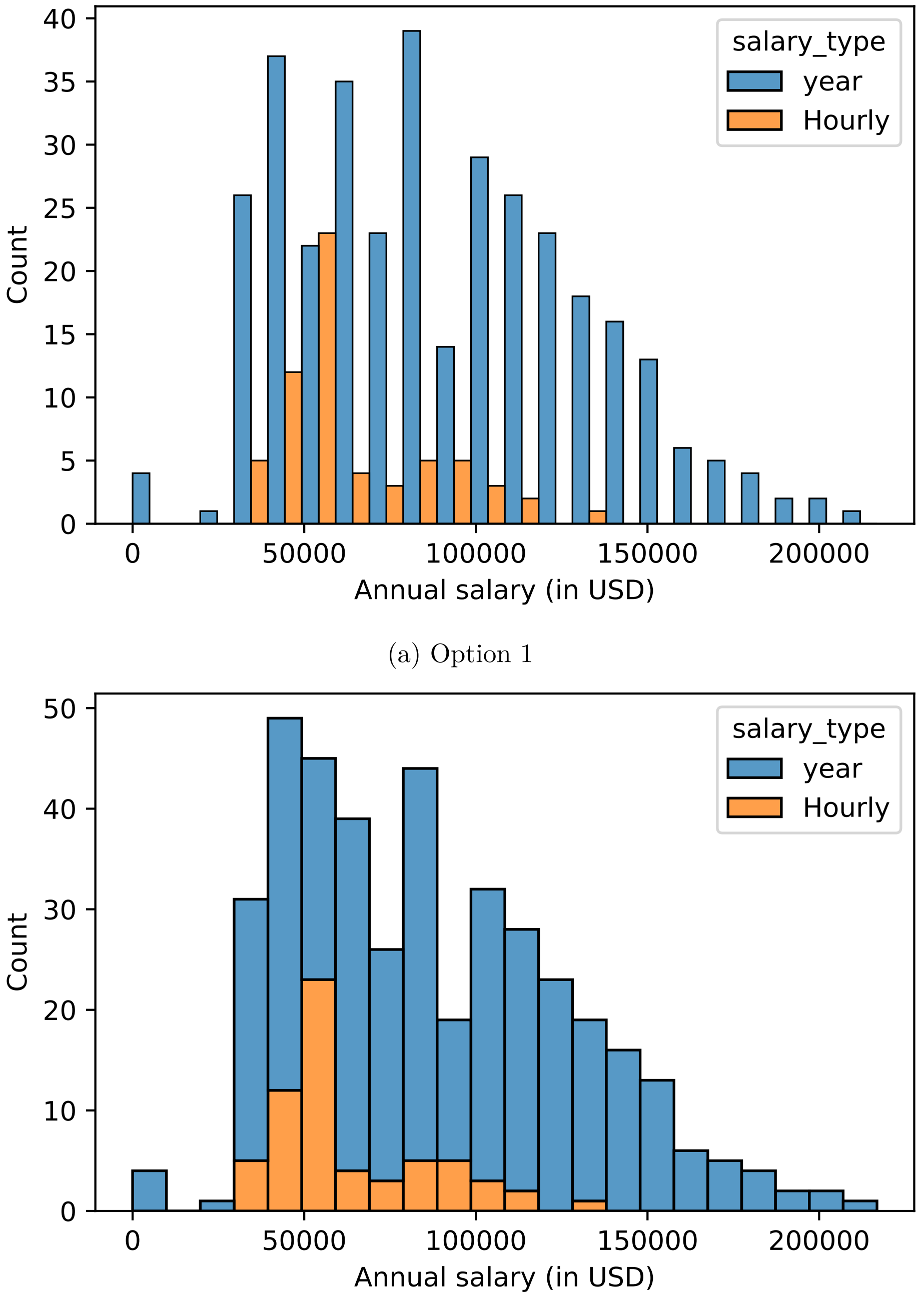
e. The `salary_type` variable is numerical.

f. The `annual_salary` variable is numerical.

g. The `performance_rating` variable is categorical and ordinal.

Question 2

Figure 1a and Figure 1b show the distributions of annual salaries of hourly and salaried workers. The two figures show the same data, with the facets organized across rows and across columns. Which of the two figures is better for comparing the median annual salaries of hourly and salaried workers. Explain your reasoning.



(a) Option 1

(b) Option 2

Figure 1: Distribution of annual salaries of Blizzard employees

Question 3

Suppose your teammate wrote the following code as part of their analysis of the data.

```
blizzard_summary = blizzard_salary.groupby('salary_type').agg(
    mean_annual_salary=('annual_salary', 'mean'),
    median_annual_salary=('annual_salary', 'median')
).reset_index()

print(blizzard_summary)
```

They then printed out the results shown below. Unfortunately one of the numbers got erased from the printout. It's indicated with _____ below.

```
  salary_type    mean_annual_salary    median_annual_salary
  Hourly         63003.                54246.
  Salaried       90183.                _____
```

Which of the following is the best estimate for that erased value?

a. 30,000

b. 50,000

c. 80,000

d. 100,000

Question 4

Which distribution of annual salaries has a higher standard deviation?

a. Hourly workers

b. Salaried workers

c. Roughly the same

Question 5

Which of the following alternate plots would also be useful for visualizing the distributions of annual salaries of hourly and salaried workers? Choose all that apply.

a. Box plots

b. Density plots

c. Pie charts

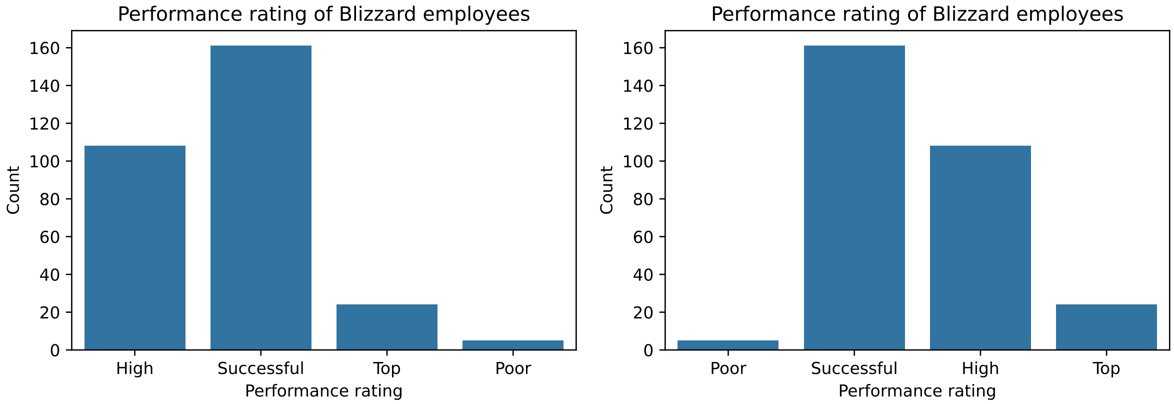
d. Waffle charts

e. Histograms

f. Scatterplots

Questions 6 and 7

Suppose you made the bar plot shown in Figure 2a to visualize the distribution of `performance_rating` and your teammate made the bar plot shown in Figure 2b.



(a) Option 1

(b) Option 2

Figure 2: Distribution of performance rating

You made your bar plot without transforming the data in any way, while your friend did first transform the data with code like the following:

```
blizzard_salary['performance_rating'] = pd._(1)_(blizzard_salary['performance_rating'], categ
```

Question 6: What goes in the blank (1)?

a. `.sort_values()`

b. `.Categorical()`

c. `.groupby()`

d. `.fillna()`

Question 7: What goes in the blank (2)?

a. `"Poor", "Successful", "High", "Top"`

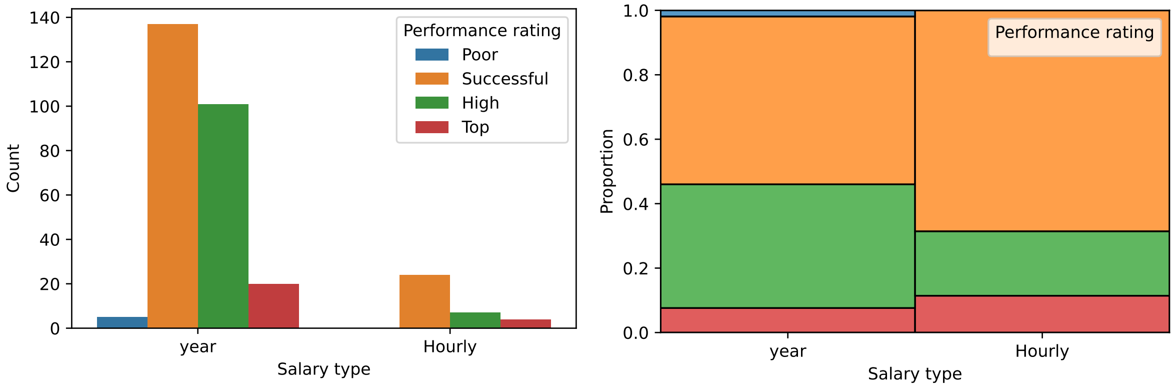
b. `"Successful", "High", "Top"`

c. `"Top", "High", "Successful", "Poor"`

d. `Poor, Successful, High, Top`

Questions 8 - 10

Finally, another teammate creates the following two plots.



(a) Option 1

(b) Option 2

Figure 3: Distribution of salary type by performance rating

Question 8: Your teammate asks you for help deciding which one to use in the final report for visualizing the relationship between performance rating and salary type. In 1-3 sentences, can you help them make a decision, justify your choice, and write the narrative that should go with the plot?

Question 9: A friend with a keen eye points out that the number of observations in Figure 3a seems lower than the total number of observations in `blizzard_salary`. What might be going on here? Explain your reasoning.

Question 10: Below are the proportions of performance ratings for hourly and salaried workers. Recreate the plot in Figure 3b, then interpret how the results from the table and within the plot relate to each other.

```
performance_rating  Poor  Successful      High       Top
salary_type
Hourly               0.0    0.149068  0.064815  0.166667
year                 1.0    0.850932  0.935185  0.833333
```

Questions 11 and 12

The table below shows the distribution of `salary_type` and `performance_rating`.

```
     percent_incr salary_type  annual_salary performance_rating
226           0.0        year        80000.0               Poor
245           3.0        year        83000.0               Poor
340           0.0        year       116000.0               Poor
391           0.0        year       135219.0               Poor
415           0.0        year       147500.0               Poor
```

The pipeline below produces a data frame with a fewer number of rows than `blizzard_salary`.

```
filtered_df = blizzard_salary[(blizzard_salary['salary_type'] _(1)_ "Hourly") _(2)_ (blizzard
filtered_df = filtered_df._(3)_(by='annual_salary')
print(filtered_df)
```

```
     percent_incr salary_type  annual_salary performance_rating
226           0.0        year        80000.0               Poor
245           3.0        year        83000.0               Poor
340           0.0        year       116000.0               Poor
391           0.0        year       135219.0               Poor
415           0.0        year       147500.0               Poor
```

Question 11: Which of the following goes in blanks (1) and (2)?

	(1)	(2)
a.	`!=`	`|`
b.	`==`	`&`
c.	`!=`	`&`
d.	`==`	`|`

Question 12: Which function or functions go into blank (3)?

a. `.sort_values()`

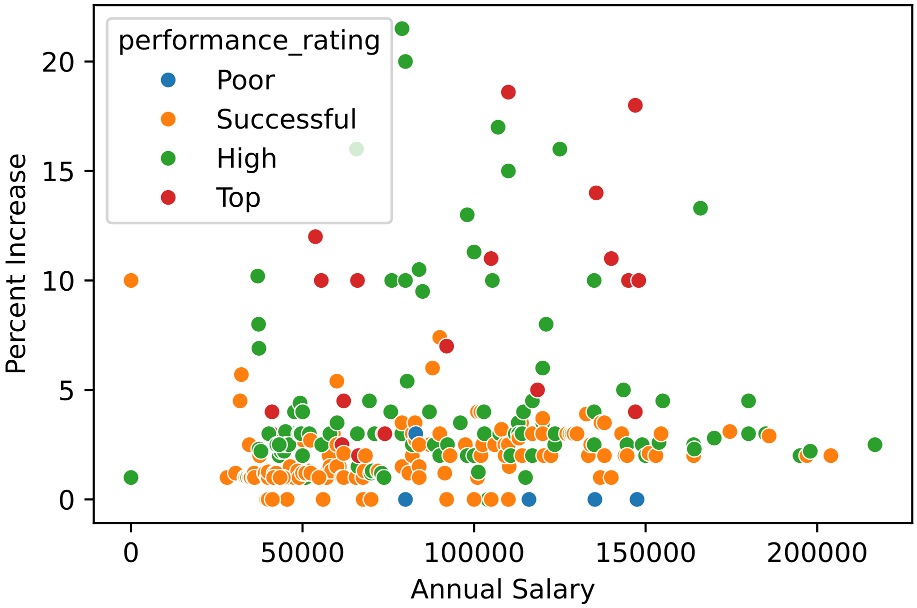
b. `.assign()`

c. `.groupby()`

d. `.aggregate()`

Question 13

You're reviewing another team's work and they made the following visualization:



And they wrote the following interpretation for the relationship between annual salary and percent increase for Top performers:

> The relationship is positive, having a higher salary results in a higher percent increase. There is one clear outlier.

Which of the following is/are the most accurate and helpful) peer review note for this interpretation. Choose all that apply.

a. The interpretation is complete and perfect, no changes needed!

b. The interpretation doesn't mention the direction of the relationship.

c. The interpretation doesn't mention the form of the relationship, which is linear.

d. The interpretation doesn't mention the strength of the relationship, which is somewhat strong.

e. There isn't a clear outlier in the plot. If any points stand out as potential outliers, more guidance should be given to the reader to identify them (e.g., salary and/or percent increase amount).

f. The interpretation is causal – we don't know if the cause of the high percent increase is higher annual salary based on observational data. The causal direction might be the other way around, or there may be other factors contributing to the apparent relationship.

Question 14

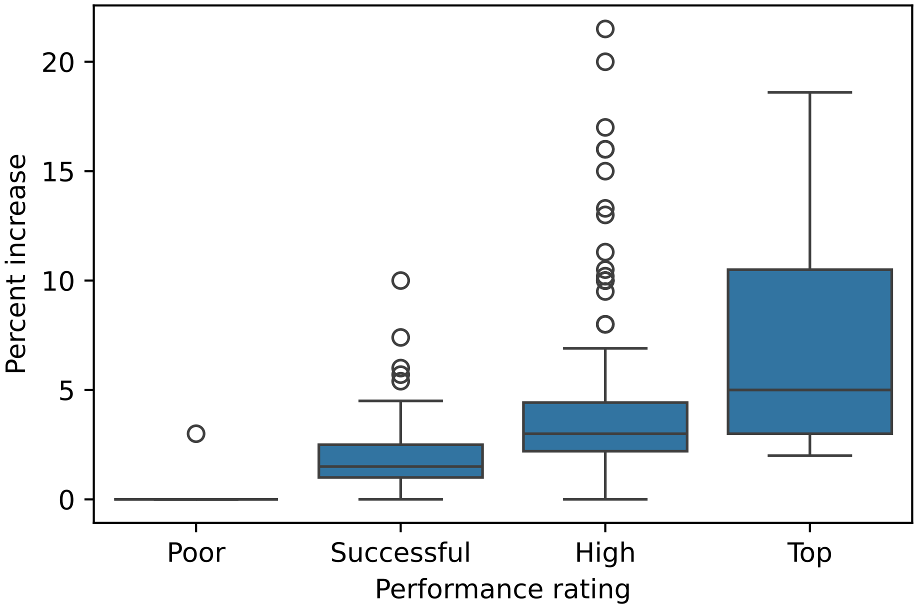
Below is some code and its output.

~~~
```{python}
# label=plot blizzard

sns.boxplot(data=blizzard_salary, x='performance_rating', y='percent_incr')
plt.xlabel('Performance rating')
plt.ylabel('Percent increase')
plt.show()

```
~~~

```
Text(0, 0.5, 'Percent increase')
```



Part 1: List at least 3 things that should be fixed or improved in the code.

Part 2: How could we show missing values in this plot?

Question 15

You're working on a data analysis on salaries of Blizzard employees in a Quarto document in a project version controlled by Git. You create a plot and write up a paragraph describing any patterns in it. Then, your teammate says "render, commit, and push".

Part 1: What do they mean by each of these three steps. In 1-2 sentences for each, explain in your own words what they mean.

1. Render:

2. Commit:

3. Push:

Part 2: Your teammate is getting impatient and they interrupt you after you rendered and committed and say "I still can't see your changes in our shared GitHub repo when I look at it in my web browser." Which of the following answers is the most accurate?

a. I rendered my document, you should be seeing my changes on GitHub when you look at it in your web browser.

b. I committed my changes, you should be seeing my changes on GitHub when you look at it in your web browser.

c. I didn't yet push my changes, it's expected that you are not seeing them on GitHub when you look at it in your web browser. Wait until I push, and check again.

d. You need to pull to see my changes on GitHub in the web browser.

Bonus

Pick a concept we introduced in class so far that you've been struggling with and explain it in your own words.