UDACITY

< Return to Classroom

Write a Data Science Blog Post

REVIEW

HISTORY

Meets Specifications

Excellent work on the analysis of women in the field of data science. You have done an amazing job on this project so far. I can see that you have written the code professionally and Your writeups about comparing salary, education, and age of men and women in the field are excellent. Congratulations on successfully completing the project.
👏🏻
Congratulations again and stay safe

Here are more reading about writing a blogpost
• How to Write an Awesome Blog Post in 5 Steps
• How to Write a Blog Post: A Step-by-Step Guide
• 7 Proven Tips to Create Blog Posts That Convert Like Crazy
• How to Write a Blog Post in 2021: The Ultimate Guide

Code Functionality and Readability

✓

Code has easy-to-follow logical structure. The code uses comments effectively and/or Notebook Markdown cells correctly. The steps of the data science process (gather, assess, clean, analyze, model, visualize) are clearly identified with comments or Markdown cells, as well. The naming for variables and functions should be according to PEP8 style guide.

The Code has easy-to-follow logical structure. Nice work using comments in the code effectively
More on adding comments to python code below:
• Python Comments
• Commenting Python Code
• Writing Comments in Python (Guide)

- How To Write Comments in Python 3
- Markdowntext

```
[9]: i = 0 # set increment number

# loop through each row of the schema table
for _, row in schema.iterrows():
    # we're just interested in the question number with certain regex
    num = re.findall(r"\d+(?=_)", row['question_num'])
    # check if current question number match the increment number
    if row['question_num'].replace('Q', '') == str(i):
        print(row['question_num'], row['question'], sep='\t')
        print()
        i += 1

    # if it doesn't match but num variable catches the regex pattern
    elif num:
        # check if the catched regex pattern matches the increment number
        if num[0] == str(i):
            print(row['question_num'], row['question'], sep='\t')
            print()
            i += 1
```


All the project code is contained in a Jupyter notebook, which demonstrates successful execution and output of the code.

Your code is contained inside a jupyter notebook and it demonstrates successful execution.

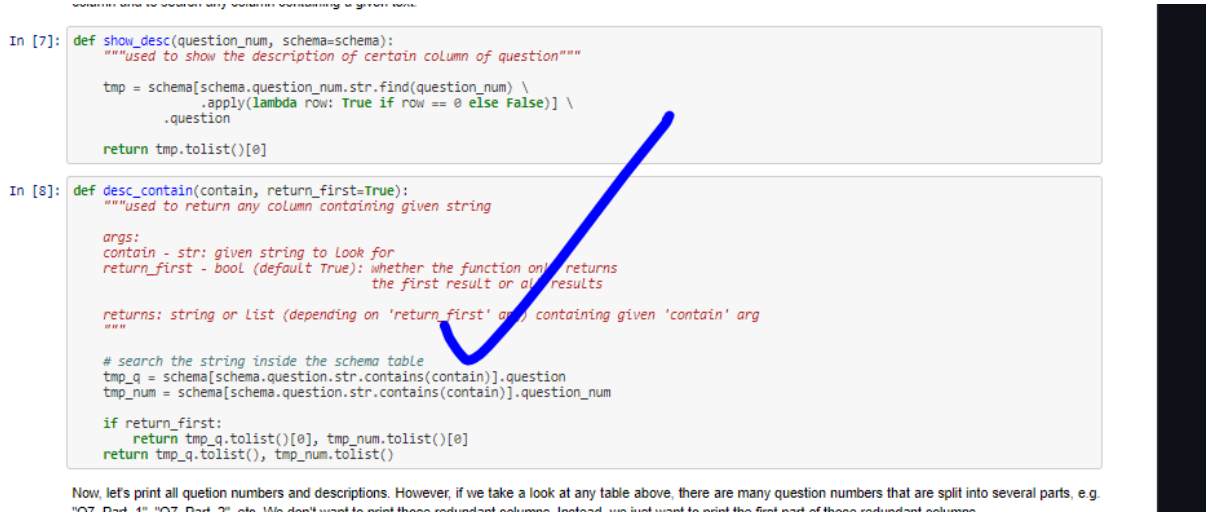
Code is well documented and uses functions and classes as necessary. All functions include document strings. DRY principles are implemented.

Very good you have used the DRY principle and have also used docstrings inside the function
You may check some links below to know more about python code documentation:

- Docstrings
- PEP257
- Documenting python code

Tips to write functions:

- Python - Functions
- Video: Introduction to Python Functions
- What are Functions?



```
In [7]: def show_desc(question_num, schema=schema):
            """used to show the description of certain column of question"""

            tmp = schema[schema.question_num.str.find(question_num) \
                         .apply(lambda row: True if row == 0 else False)] \
                  .question

            return tmp.tolist()[0]

In [8]: def desc_contain(contain, return_first=True):
            """used to return any column containing given string

            args:
            contain - str: given string to look for
            return_first - bool (default True): whether the function only returns
                                               the first result or all results

            returns: string or list (depending on 'return_first' arg) containing given 'contain' arg
            """

            # search the string inside the schema table
            tmp_q = schema[schema.question.str.contains(contain)].question
            tmp_num = schema[schema.question.str.contains(contain)].question_num

            if return_first:
                return tmp_q.tolist()[0], tmp_num.tolist()[0]
            return tmp_q.tolist(), tmp_num.tolist()
```

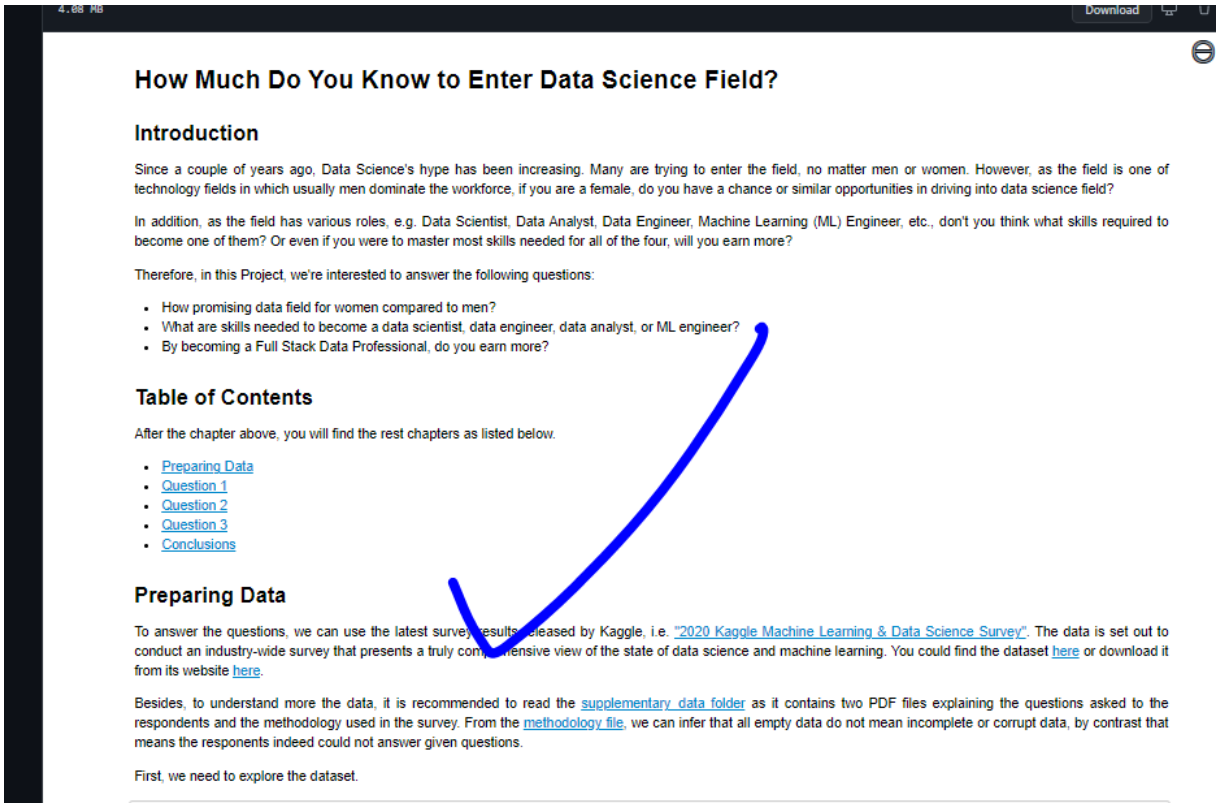
Now, let's print all quetion numbers and descriptions. However, if we take a look at any table above, there are many question numbers that are split into several parts, e.g.

Data

Project follows the CRISP-DM process outlined for questions through communication. This can be done in the README or the notebook. If a question does not require machine learning, descriptive or inferential statistics should be used to create a compelling answer to a particular question.

Excellent you have used the CRISP-DM process in the notebook
More on CRISP-DM process at below:

- CIRSP-DM to predict car prices
- CRISP-DM methodology
- What is CRISP-DM
- How to perform Data Analysis using the CRISP-DM approach?



How Much Do You Know to Enter Data Science Field?

Introduction

Since a couple of years ago, Data Science's hype has been increasing. Many are trying to enter the field, no matter men or women. However, as the field is one of technology fields in which usually men dominate the workforce, if you are a female, do you have a chance or similar opportunities in driving into data science field?

In addition, as the field has various roles, e.g. Data Scientist, Data Analyst, Data Engineer, Machine Learning (ML) Engineer, etc., don't you think what skills required to become one of them? Or even if you were to master most skills needed for all of the four, will you earn more?

Therefore, in this Project, we're interested to answer the following questions:

- How promising data field for women compared to men?
- What are skills needed to become a data scientist, data engineer, data analyst, or ML engineer?
- By becoming a Full Stack Data Professional, do you earn more?

Table of Contents

After the chapter above, you will find the rest chapters as listed below.

- Preparing Data
- Question 1
- Question 2
- Question 3
- Conclusions

Preparing Data

To answer the questions, we can use the latest survey results released by Kaggle, i.e. "2020 Kaggle Machine Learning & Data Science Survey". The data is set out to conduct an industry-wide survey that presents a truly comprehensive view of the state of data science and machine learning. You could find the dataset here or download it from its website here.

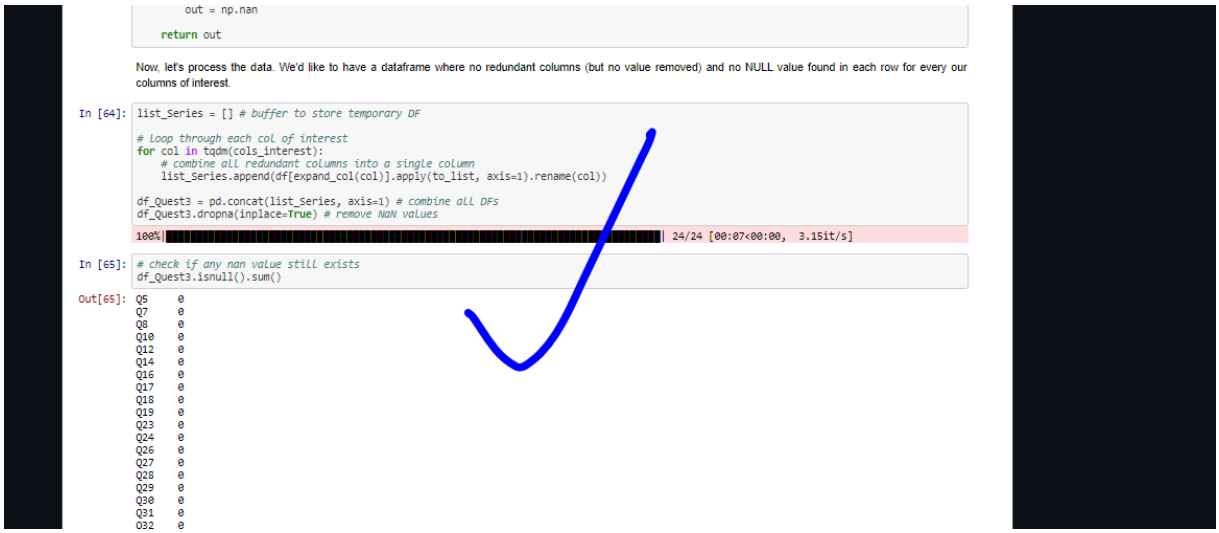
Besides, to understand more the data, it is recommended to read the supplementary data folder as it contains two PDF files explaining the questions asked to the respondents and the methodology used in the survey. From the methodology file, we can infer that all empty data do not mean incomplete or corrupt data, by contrast that means the responents indeed could not answer given questions.

First, we need to explore the dataset.

Categorical variables are handled appropriately for machine learning models (if models are created). Missing values are also handled appropriately for both descriptive and ML techniques. Document why a particular approach was used, and why it was appropriate for a particular situation.

Very good, you have handled categorical variables, and you have discussed how you handled the missing values
More on handling missing data at below:

- The prevention and handling of the missing data
- How to Handle Missing Data
- Dealing with Missing Data
- The best way to handle missing data



```
        out = np.nan

    return out
```

Now, let's process the data. We'd like to have a dataframe where no redundant columns (but no value removed) and no NULL value found in each row for every our columns of interest.

In [64]:
```
list_Series = [] # buffer to store temporary DF

# Loop through each col of interest
for col in tqdm(cols_interest):
    # combine all redundant columns into a single column
    list_Series.append(df[expand_col(col)].apply(to_list, axis=1).rename(col))

df_Quest3 = pd.concat(list_Series, axis=1) # combine all DFs
df_Quest3.dropna(inplace=True) # remove NaN values
```

```
100%|██████████| 24/24 [00:07<00:00,  3.15it/s]
```

In [65]:
```
# check if any nan value still exists
df_Quest3.isnull().sum()
```

Out[65]:
```
Q5     0
Q7     0
Q8     0
Q10    0
Q12    0
Q14    0
Q16    0
Q17    0
Q18    0
Q19    0
Q23    0
Q24    0
Q26    0
Q27    0
Q28    0
Q29    0
Q30    0
Q31    0
Q32    0
```

Analysis, Modeling, Visualization

✓

In the Jupyter Notebook, there are between 3-5 questions asked, related to the business or real-world context of the data. Each question is answered with appropriate visualization, table, or statistic.

Nice job with your questions! These are interesting questions, and you did a great job of showing clean visualizations for each. You didn't go overboard on providing too much information in any single visualization, but presented your ideas in a clear succinct way

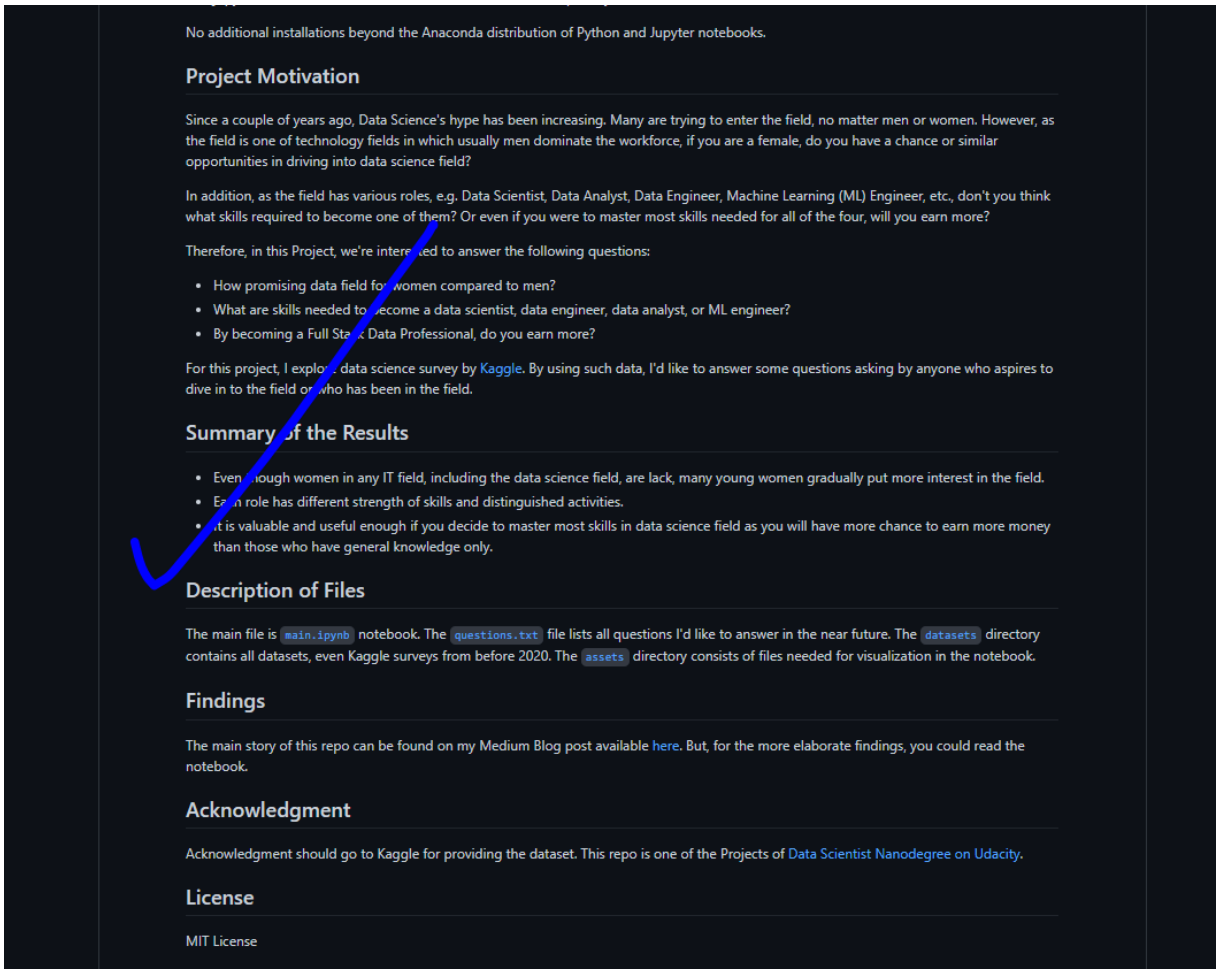
Github Repository

✓

Student must have a Github repository of their project. The repository must have a README.md file that communicates the libraries used, the motivation for the project, the files in the repository with a small description of each, a summary of the results of the analysis, and necessary acknowledgements. Students should not use another student's code to complete the project, but they may use other references on the web including StackOverflow and Kaggle to complete the project.

Nice job posting your code to Github. Your README looks great! You have installation, project motivation, file descriptions, results, and acknowledgments. I would, however, add a visual or two.

More on Readme here
• Here are few great examples
• How do you put Images on the README.md file?
• What Media Sources Data Scientists using the Most? Analysis of the 2019 Kaggle ML & DS Survey
• Top 7 BENEFITS OF USING GITHUB



No additional installations beyond the Anaconda distribution of Python and Jupyter notebooks.

Project Motivation

Since a couple of years ago, Data Science's hype has been increasing. Many are trying to enter the field, no matter men or women. However, as the field is one of technology fields in which usually men dominate the workforce, if you are a female, do you have a chance or similar opportunities in driving into data science field?

In addition, as the field has various roles, e.g. Data Scientist, Data Analyst, Data Engineer, Machine Learning (ML) Engineer, etc., don't you think what skills required to become one of them? Or even if you were to master most skills needed for all of the four, will you earn more?

Therefore, in this Project, we're interested to answer the following questions:

- How promising data field for women compared to men?
- What are skills needed to become a data scientist, data engineer, data analyst, or ML engineer?
- By becoming a Full Stack Data Professional, do you earn more?

For this project, I explore data science survey by Kaggle. By using such data, I'd like to answer some questions asking by anyone who aspires to dive in to the field or who has been in the field.

Summary of the Results

- Even though women in any IT field, including the data science field, are lack, many young women gradually put more interest in the field.
- Each role has different strength of skills and distinguished activities.
- It is valuable and useful enough if you decide to master most skills in data science field as you will have more chance to earn more money than those who have general knowledge only.

Description of Files

The main file is `main.ipynb` notebook. The `questions.txt` file lists all questions I'd like to answer in the near future. The `datasets` directory contains all datasets, even Kaggle surveys from before 2020. The `assets` directory consists of files needed for visualization in the notebook.

Findings

The main story of this repo can be found on my Medium Blog post available here. But, for the more elaborate findings, you could read the notebook.

Acknowledgment

Acknowledgment should go to Kaggle for providing the dataset. This repo is one of the Projects of Data Scientist Nanodegree on Udacity.

License

MIT License

Blog Post

✓

Student must have a blog post on a platform of their own choice (can be on their website, a Medium post or Github blog post). The post should not dive into technical details or difficulties of the analysis - this should be saved for Github. The post should be understandable for non-technical people from many fields.

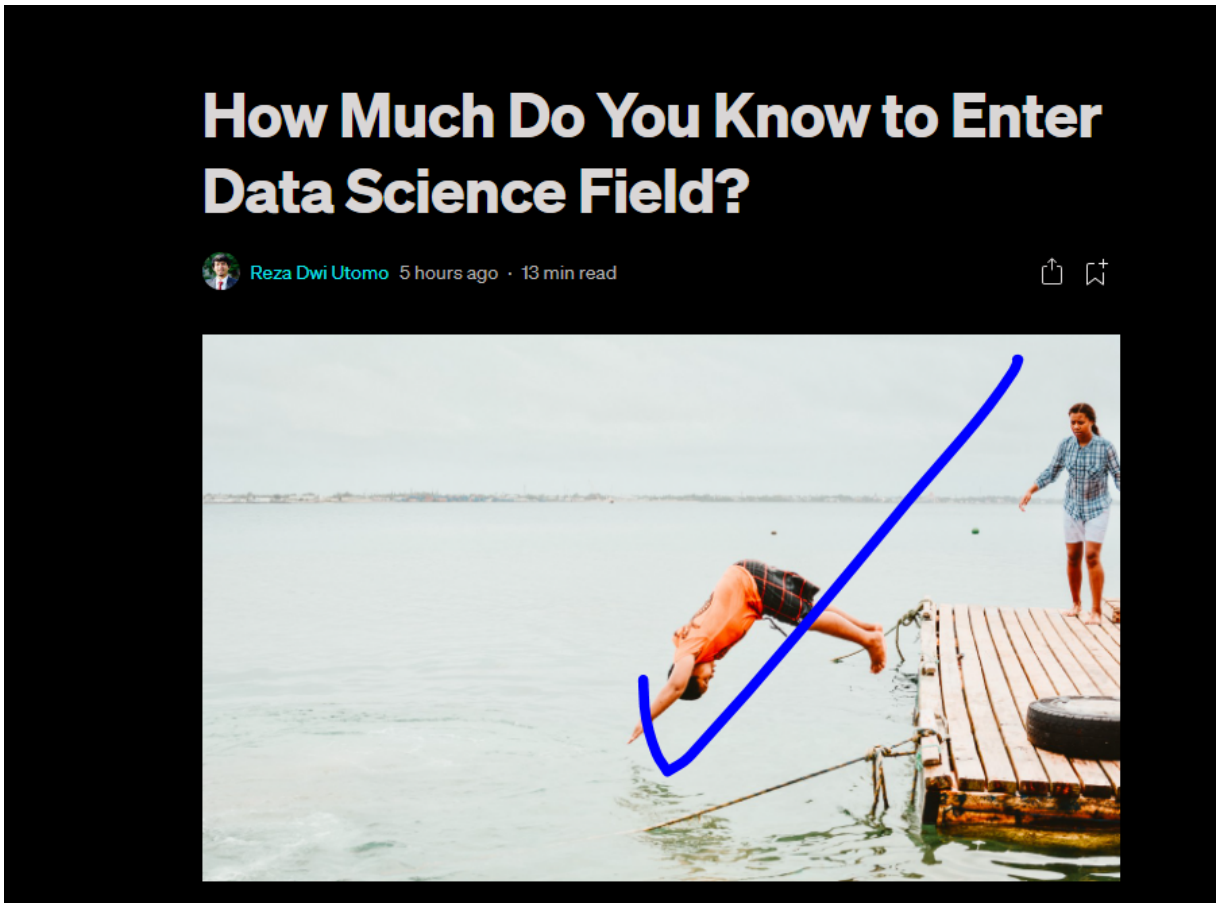
Blogpost is nicely written for non technical audience

✓

Student must have a title and image to draw readers to their post.

Excellent choice of title and image at the top of the blog. Image is an excellent way to draw readers
More on choosing the blogpost image at below:
• How To Create Blog Title Images That Attract Readers
• 11 Best Practices for Including Images in Your Blog Posts
• How to Select the Perfect Image for Your Next Blog Post

• Blog Image 101 - The most effective image tips to help your post stand out



✓

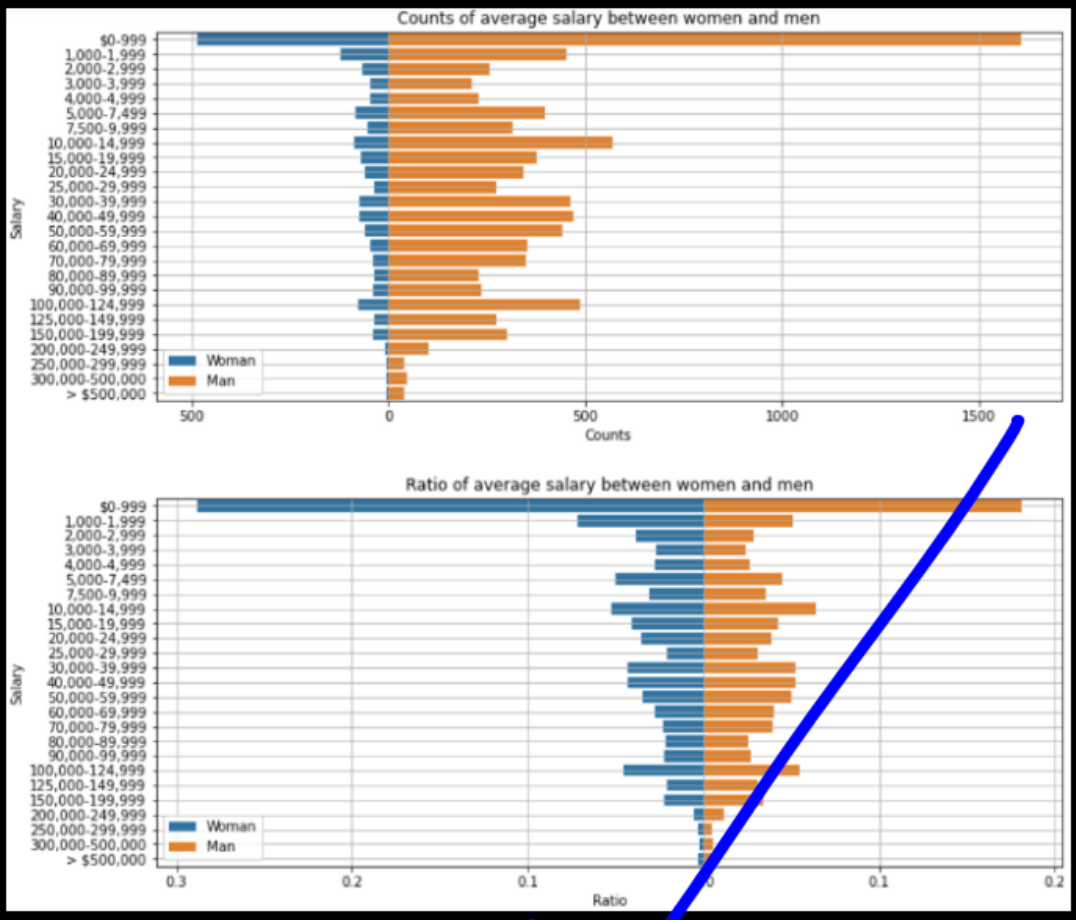
There are no long, ongoing blocks of text without line breaks or images for separation anywhere in the post.

You have done well to avoid long paragraphs

✓

Each question is clearly stated and each answer includes a clear visual, table, or statistic.

Very well done, each question is answered with a clear visual, table or statistics that provide how the data supports the hypothesis.
Modern charting:
• plot.ly: For creating interactive visuals
• seaborn: statistical data visualization. Allows user to create better visual without much coding effort



From both charts above, we can see, again, men always dominate in terms of numbers. But, if focus on the ratio chart, there is a large proportion for the least salary range in women data. This proportion should be for those women with minimum experience since usually less experienced employees earn less than the more experienced ones. However, to prove this hypothesis, let's explore this proportion where the gender is woman and salary range is between 0 and 999 USD.

PROJECT LINK

RETURN TO PATH