Student Information

Name: **[**Fill Your Name**]**

Student ID: 1155000000

GitHub ID: [**GitHub ID**]

Instructions

1. First, click the **Make a copy** button in the Google Drive popup window. You may need to scroll some words from the call-to-action button to the top of the popup window to find it.

2. Rename Notebook file name prefixes from the DSMLC-L01 Jupyter Notebook **to your Student ID**. You don't need to explain all details, but do explain **minimal comments** explaining your code and code-through. **You just to work 50% of your grade.**

3. Complete the Notebook tasks. The assignment contains a **QUESTION** and **ANSWER** section. Write the specifications of the answers in the **ANSWER** section.

   - You are allowed to use any libraries, but **DO NOT** hard-code or fix the data or the test set values since they need to be generalized.

4. Share access to your Notebook by setting to **the sole listener**. **This just to work 50% of your grade.**

   - Sensitive information **are data visualizations**. Refer to some resources and the Data Wrangling lecture for inspiration and ideas.

   - Click the blue button at the top-right of your screen, then **Share** and change **General access** to **Anyone with the link**. Change the access from **Viewer** to **Editor**. Copy the link and share it with **only** the instructor and TA by email. Alternatively, you can share with the instructor and TA by their Google Drive accounts (their @connect.ust.hk accounts).

   - Check if the link you copied and pasted into Canvas works properly. Make sure you can open the link in a new private browser window to see if it works properly. Do not share your Google Drive with your classmates.

5. Submit your work by **Friday, Week 10 (Nov. 4th 11:59 pm, Thursday)**

Data Mining Lab 1 Solution

Part 1

Titanic Machine Learning from Disaster (the DSMLC01-Lab1 Kaggle Task)

> **QUESTION:** Write a function, `"def dataset_1(x):"`, to return answer set 1.

Table of Contents

1. Data Source
2. Data Preprocessing
   - Exercise 2
3. Data Transformation
   - Exercise 3 (pandas)
   - Exercise 4
   - Exercise 5 (table frame)
4. Data Mining using Pandas
   - Exercise 6 (pandas)
   - Exercise 7 (table frame)
   - Exercise 8 (table frame)
   - Exercise 9 (table frame)
   - Exercise 10 (table frame)
   - Exercise 11 (table frame)
   - Exercise 12 (table frame)
   - Exercise 13 (table frame)
   - Exercise 14 (table frame)
   - Exercise 15 (table frame)
   - Exercise 16 (table frame)
   - Exercise 17 (table frame)

1. The Data

```
# Kaggle Titanic
# https://www.kaggle.com/c/titanic/data
# Task: Binary classification
# https://www.kaggle.com/c/titanic/data?select=train.csv
# https://www.kaggle.com/c/titanic/data?select=test.csv
# https://www.kaggle.com/c/titanic/data?select=gender_submission.csv
# https://www.kaggle.com/c/titanic/data?select=titanic.zip
# https://www.kaggle.com/c/titanic/data?select=titanic3.xls
# https://www.kaggle.com/c/titanic/data?select=titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.zip/titanic_names/titanic_names.txt
# https://www.kaggle.com/c/titanic/data?select=titanic_names.zip/titanic_names/
```

>>> Exercise 12 (take home):

Please try to plot the Dataset in 3 and 2 and the most use 3-D plot. Use at least 5 different single-colors palettes to check how your data and features are distributed.

(Hint: you can refer to Seaborn to the documentation)

```
# Plot the 3-D plot of the dataset
%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(111, projection='3d')

# Plot the training points
for name, label in [('Setosa', 0),
                    ('Versicolour', 1),
                    ('Virginica', 2)]:
    ax.text3D(X_reduced[y == label, 0].mean(),
              X_reduced[y == label, 1].mean(),
              X_reduced[y == label, 2].mean,
              name,
              bbox=dict(alpha=.5, edgecolor='w', facecolor='w'),
              horizontal_alignment='center',
              bbox=False,
              color=np.random.rand(3,1,3),
              size=20, zorder=3.)
# [::-1] means reverse the order
y = np.choose(y, [1, 2, 0])
ax.scatter(X_reduced[:, 0], X_reduced[:, 1], X_reduced[:, 2],
           c=y, cmap=cmap_bold,
           edgecolor='k', s=40)

ax.set_title("3-D Iris Dataset")
ax.set_axis_off()
ax.dist = 12
fig.show()
```

>>> Exercise 13 (take home):

If you plot a histogram of the petal-length feature, I would encourage you try to break and see plots by actives this.

```
# Create a histogram of the petal length
# Create a histogram of the petal length
# Create a histogram of the petal length
ax = plot.hist(iris.petal_length, bins=20, color='c')
plot.xlabel('Petal Length')
plot.ylabel('Frequency')
plot.show()

# Create a histogram of the petal length
# Create a histogram of the petal length
# Create a histogram of the petal length
ax = plot.hist(iris.petal_length, bins=20, color='c')
plot.xlabel('Petal Length')
plot.ylabel('Frequency')
plot.show()
```

>>> Exercise 14 (take home):

The next step is to explore the data. It is actually, and it is extraordinarily interesting to both computer and visualize. Can you efficiently explore the data? Please use at least 5 different ways to explore.

```
# Create the
# Create the histogram of the data
# Create the histogram of the data
# Create the histogram of the data
ax = plot.hist(iris.petal_length, bins=20, color='c')
plot.xlabel('Petal Length')
plot.ylabel('Frequency')
plot.show()
# Create the histogram of the data
# Create the histogram of the data
# Create the histogram of the data
ax = plot.hist(iris.petal_length, bins=20, color='c')
plot.xlabel('Petal Length')
plot.ylabel('Frequency')
plot.show()
```

>>> Exercise 15 (take home):

Additionally, we can attempt to find the frame or the **Gini(G)**, by frequency instead of a alphabetical order. The way the classification is more meaningful and you will be able to observe the so-called Gini or Left-right bias with the error which will appear a lot from the coding and other observations. Please

```
# Create the
# Create the histogram of the data
# Create the histogram of the data
# Create the histogram of the data
ax = plot.hist(iris.petal_length, bins=20, color='c')
plot.xlabel('Petal Length')
plot.ylabel('Frequency')
plot.show()
# Create the histogram of the data
# Create the histogram of the data
# Create the histogram of the data
ax = plot.hist(iris.petal_length, bins=20, color='c')
plot.xlabel('Petal Length')
plot.ylabel('Frequency')
plot.show()
```

>>> Exercise 16 (take home):

Try to generalize the transformation using the `category_name` column instead. Does it work?

```
# Create
# Create
# Create
```

```
# Answer here:
# Exercise 7 (5 min):
# Notice that the 'JOB' attribute has hashtags the maximum value for a 'JOB' is possible to summarize this concept of hash counting.
# Find would you go about doing that? (Hint: look at code above for clues)

df[df['JOB'].str.contains('#')]
df['JOB'].value_counts.sort_values.reset_index.head(10)
print(df[df['JOB'].isnull()])
df['JOB'].fillna('unknown', inplace=True)
df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                 text  text  company  company.dup
1  Sorbonne is my top 1k hit so big to have i...     1  Amazon     1
2                                  Good idea, Excellent idea.     1  Amazon     1
3                                                Where the hell is     1  Amazon     1
4  There's changes for commutators training ideas...     1  Amazon     1
5                                                   Thanks for the suggestion     1  Amazon     1
6  I think it applies the most part with the big ...     1  Amazon     1
7  If you have several dozen or several hundred c...     1  Amazon     1
8  I've found the week, you can find the ...     1  Amazon     1
9                                                   Thanks for the suggestion     1  Amazon     1
10                                                 What a waste of money and time!     1  Amazon     1
Series([], Name: JOB, dtype: int64)
Series([], Name: JOB, dtype: int64)
Empty DataFrame
Columns: [text, text, company, company.dup]
Index: []
Series([], Name: JOB, dtype: int64)
```

```
# Exercise 8 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 9 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                 text  text  company  company.dup
1  Sorbonne is my top 1k hit so big to have i...     1  Amazon     1
2                                  Good idea, Excellent idea.     1  Amazon     1
3                                                Where the hell is     1  Amazon     1
4  There's changes for commutators training ideas...     1  Amazon     1
5                                                   Thanks for the suggestion     1  Amazon     1
6  I think it applies the most part with the big ...     1  Amazon     1
7  If you have several dozen or several hundred c...     1  Amazon     1
8  I've found the week, you can find the ...     1  Amazon     1
9                                                   Thanks for the suggestion     1  Amazon     1
10                                                 What a waste of money and time!     1  Amazon     1
Series([], Name: JOB, dtype: int64)
Series([], Name: JOB, dtype: int64)
Empty DataFrame
Columns: [text, text, company, company.dup]
Index: []
Series([], Name: JOB, dtype: int64)
```

```
# Exercise 10 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 11 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                 text  text  company  company.dup
1  Sorbonne is my top 1k hit so big to have i...     1  Amazon     1
2                                  Good idea, Excellent idea.     1  Amazon     1
3                                                Where the hell is     1  Amazon     1
4  There's changes for commutators training ideas...     1  Amazon     1
5                                                   Thanks for the suggestion     1  Amazon     1
6  I think it applies the most part with the big ...     1  Amazon     1
7  If you have several dozen or several hundred c...     1  Amazon     1
8  I've found the week, you can find the ...     1  Amazon     1
9                                                   Thanks for the suggestion     1  Amazon     1
10                                                 What a waste of money and time!     1  Amazon     1
Series([], Name: JOB, dtype: int64)
Series([], Name: JOB, dtype: int64)
Empty DataFrame
Columns: [text, text, company, company.dup]
Index: []
Series([], Name: JOB, dtype: int64)
```

```
# Exercise 12 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 13 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 14 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 15 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 16 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 17 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 18 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 19 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 20 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 21 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 22 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 23 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 24 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 25 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 26 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 27 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 28 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 29 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 30 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 31 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 32 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 33 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 34 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 35 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 36 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 37 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 38 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 39 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 40 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 41 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 42 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 43 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 44 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 45 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 46 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 47 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 48 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 49 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 50 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 51 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['JOB'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 52 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['text'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 53 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .value_counts() function.

df['company'].value_counts.sort_values.reset_index.head(10)
```

```
                                                text  text
0                                   Good idea, Excellent idea.     1
1                                                 Where the hell is     1
2                                                    Thanks for the suggestion     1
3                                                    Thanks for the suggestion     1
4                                                    Thanks for the suggestion     1
5                                                    Thanks for the suggestion     1
6                                                    Thanks for the suggestion     1
7                                                    Thanks for the suggestion     1
8                                                    Thanks for the suggestion     1
9                                                    Thanks for the suggestion     1
```

```
# Exercise 54 (5 min):
# Suppose the Sorbonne is the top hit for a search for 'Sorbonne'. Assuming 'Sorbonne' is the first hit, how many times would you expect to see 'Sorbonne' in the top 10 results?
# Hint: Use the .
```

```
[1]: pickle.dump(BC_df,open(
                 'BC_df.pkl','wb'))
[2]: BC_df.groupby('Category')
     [['Length','Punctuation']].
     mean().round(2).reset_index()
[2]:  Category  Length  Punctuation
      0    ham   71.45        9.67
      1   spam  138.67       29.89
[3]: plt.bar(x=BC_df.groupby('Category')
             ['Length'].mean().round(2).index,
             height=BC_df.groupby('Category')
             ['Length'].mean().round(2).values,
             color=['lightblue','lightgreen'])
     plt.ylabel('Length')
     plt.xlabel('Category')
     plt.title('Average Length by Category')
     plt.text(x=0,y=71.45+3,s=71.45)
     plt.text(x=1,y=138.67+3,s=138.67)
     plt.show()
```


We found that spam emails are twice as long as ham emails.

```
[4]: BC_df.shape, BC_df.head(1).shape
```

```
[5]: (5572, 4),
     (1, 4)
```

```
[6]: BC_df[BC_df.Length==910]
```

```
[7]:  Category                                            Message  Length  \
      1   spam  Free entry in 2 a wkly comp to win FA Cup fina...     910
      2   spam  FREE! Text WORLD to 88070 to receive 1st ringt...     910
      3   spam  Free entry in 2 a wkly comp to win FA Cup fina...     910
      4   spam  Free entry in 2 a wkly comp to win FA Cup fina...     910
      5   spam  Free entry in 2 a wkly comp to win FA Cup fina...     910
```