LET'S GO!

It's time to book a holiday

An in-depth analysis of our favourite destinations

Introduction

THE DATA SET

· Covers 15 of the most popular holiday destination in the world
· Uses information gathered from across the top 5 travel review websites
· Actually, I made it all up
· And I am biased, because I am from Cape Town...
...But its a great exercise in data filtering and visualisation
·

What's In The Data Set?

DESTINATION COUNTRY

Top 15 destination countries, including the most visited city in each

FEEDBACK SCORES

A feedback score out of 10 for each destination, aggregated over thousands of reviews

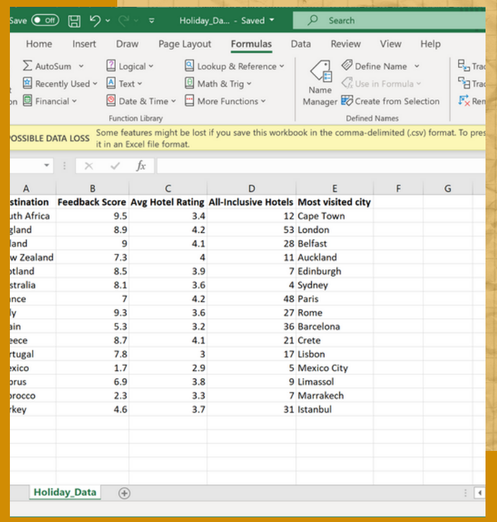
ALL-INCLUSIVE HOTELS

The number of all-inclusive hotels available in each destination

AVERAGE HOTEL RATINGS

An aggregated rating out of 5 stars for all of the hotels for each destination

Step 1: Create the data set in excel and import to VSCode, showing the number of rows and columns



Holiday_Data csv

```
1. Show how many rows and columns
   there are in the file

data.shape  # Shows that the data has
            # 15 rows and 5 columns
[150]  ✓ 0.5s
...  (15, 5)
```

VS Code script

```
import pandas as pd
data = pd.read_csv("Holiday_Data.csv")
print(data)       # Sanity check that the data looks correct
✓ 0.7s
```

	Destination	Feedback Score	Avg Hotel Rating	All-Inclusive Hotels	Most visited city
0	South Africa	9.5	3.4	12	Cape Town
1	England	8.9	4.2	53	London
2	Ireland	9.0	4.1	28	Belfast
3	New Zealand	7.3	4.0	11	Auckland
4	Scotland	8.5	3.9	7	Edinburgh
5	Australia	8.1	3.6	4	Sydney
6	France	7.0	4.2	48	Paris
7	Italy	9.3	3.6	27	Rome
8	Spain	5.3	3.2	36	Barcelona
9	Greece	8.7	4.1	21	Crete
10	Portugal	7.8	3.0	17	Lisbon
11	Mexico	1.7	2.9	5	Mexico City
12	Cyprus	6.9	3.8	9	Limassol
13	Morocco	2.3	3.3	7	Marrakech
14	Turkey	4.6	3.7	31	Istanbul

The data set has been imported correctly

STEP 2: PRINT ROWS 3 TO 8 FROM THE DATA SET

Use the iloc/loc functions in pandas

```
2. Print row 3-8 (using iloc/loc)

print(data.iloc[3:9]) # up to but not including row 9
[ ]
```

VS Code Script

```
print(data.iloc[3:9]) # up to but not including row 9
[151] ✓ 0.5s
```

```
...    Destination  Feedback Score  Avg Hotel Rating  All-Inclusive Hotels  \
3  New Zealand              7.3               4.0                    11
4     Scotland              8.5               3.9                     7
5    Australia              8.1               3.6                     4
6       France              7.0               4.2                    48
7        Italy              9.3               3.6                    27
8        Spain              5.3               3.2                    36

  Most visited city
3          Auckland
4         Edinburgh
5            Sydney
6             Paris
7              Rome
8         Barcelona
```

VS Code Output

Step 3: Find the mean number of all-inclusive hotels across all destinations

```
3. Find the mean number of all-inclusive hotels across all destinations

Mean_AllInclusive = data["All-Inclusive Hotels"].mean()
print(Mean_AllInclusive)
[153]  ✓ 0.5s
...  21.066666666666666
```

VS Code output

```
3. Find the mean number of all-inclusive hotels across all destinations

Mean_AllInclusive = data["Number of All-Inclusive Hotels"].mean()
print(Mean_AllInclusive)
[ ]
```

VS Code script

This is shown to be 21.067

```
Lowest_Score = data["Feedback Score"].min()
Lowest_Scoring_Destination = pd.read_csv("Holiday_Data.csv",index_col = "Feedback Score")
Lowest_Scoring_Destination = Lowest_Scoring_Destination.loc[[Lowest_Score]]
print(Lowest_Scoring_Destination[["Destination"]])
```

VS Code script

STEP 4. FIND THE LOWEST SCORING DESTINATION

This is shown to be Mexico, with a score of 1.7

```
                Destination
Feedback Score
1.7                  Mexico
```

VS Code output

Step 5: Find the highest scoring destination

This is shown to be South Africa, with a score of 9.5

```
                    Destination
Feedback Score
9.5                 South Africa
```

VS Code output

```
5. Find the highest scoring destination
```

```
Highest_Score = data["Feedback Score"].max()
Highest_Scoring_Destination = pd.read_csv("Holiday_Data.csv",index_col = "Feedback Score")
Highest_Scoring_Destination = Highest_Scoring_Destination.loc[[Highest_Score]]
print(Highest_Scoring_Destination[["Destination"]])
```

VS Code script

```
6. Find the destinations where there are more than 9 all-inclusive hotels
```

```
AllInc_Above9_Filter = data["All-Inclusive Hotels"] > 9  # Returns booleans
AllInc_Above9 = data[AllInc_Above9_Filter] # Uses booleans to filter data for True
print(AllInc_Above9)
```

[ ]

VS Code output

	Destination	Feedback Score	Avg Hotel Rating	All-Inclusive Hotels
0	South Africa	9.5	3.4	12
1	England	8.9	4.2	53
2	Ireland	9.0	4.1	28
3	New Zealand	7.3	4.0	11
6	France	7.0	4.2	48
7	Italy	9.3	3.6	27
8	Spain	5.3	3.2	36
9	Greece	8.7	4.1	21
10	Portugal	7.8	3.0	17
14	Turkey	4.6	3.7	31

	Most visited city
0	Cape Town
1	London
2	Belfast
3	Auckland
6	Paris
7	Rome
8	Barcelona
9	Crete
10	Lisbon
14	Istanbul

VS Code script

STEP 6: FIND ALL THE DESTINATIONS WHERE THERE ARE MORE THAN 9 ALL-INCLUSIVE HOTELS

10 out of the 15 destinations satisfied this criterion

Step 7: Filter the data by score above 8

```
7. Filter the data by score above 8


Score_Above_8_Filter = data["Feedback Score"] > 8 # Returns booleans
Score_Above_8 = data[Score_Above_8_Filter]
print(Score_Above_8)
```

VSCode Script

```
    Destination  Feedback Score  Avg Hotel Rating  All-Inclusive Hotels  \
0  South Africa             9.5               3.4                    12
1       England             8.9               4.2                    53
2       Ireland             9.0               4.1                    28
4      Scotland             8.5               3.9                     7
5     Australia             8.1               3.6                     4
7         Italy             9.3               3.6                    27
9        Greece             8.7               4.1                    21

  Most visited city
0        Cape Town
1           London
2          Belfast
4         Edinburgh
5           Sydney
7             Rome
9            Crete
```

VS Code Output

There are 6 destinations that fulfill this criterion

STEP 8: FILTER THE DATA BY SCORE BELOW 2

Only 1 destination had a score below 2 - Mexico.

Further investigation into this result is needed to validate its inclusion in the data set

```
8. Filter the data score below 2



Score_Below_2_Filter = data["Feedback Score"] < 2 # Returns booleans
Score_Below_2 = data[Score_Below_2_Filter]
print(Score_Below_2)
```

VS Code Script

```
   Destination  Feedback Score  Avg Hotel Rating  All-Inclusive Hotels  \
11      Mexico             1.7               2.9                     5

   Most visited city
11       Mexico City
```

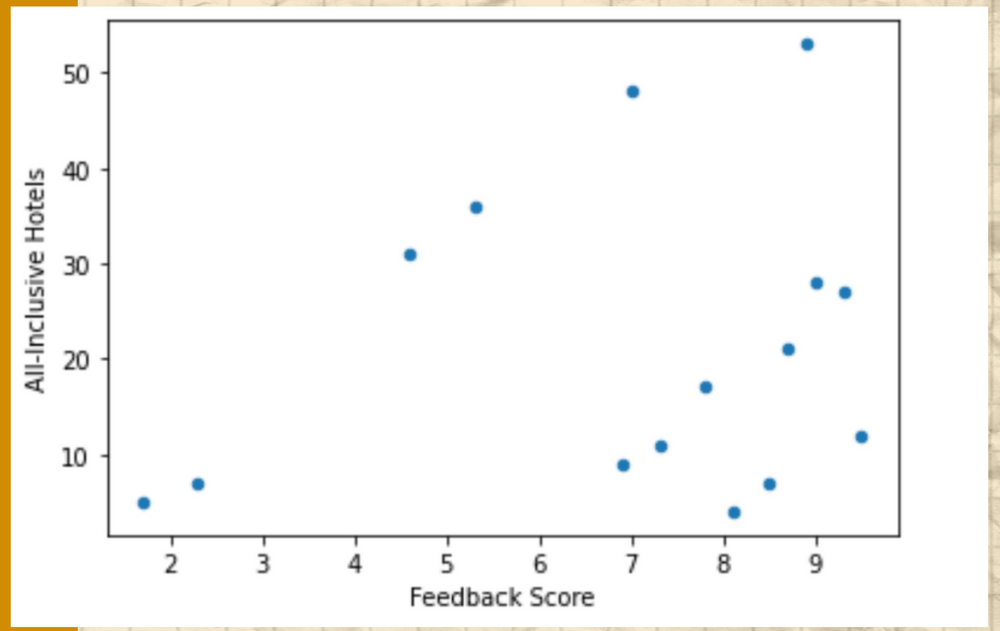
VS Code Output

What makes a holiday destination popular?

Surprisingly, there appears to be no correlation between a destination's review score and the number of all-inclusive hotels available.

VS Code script to create a scatter plot

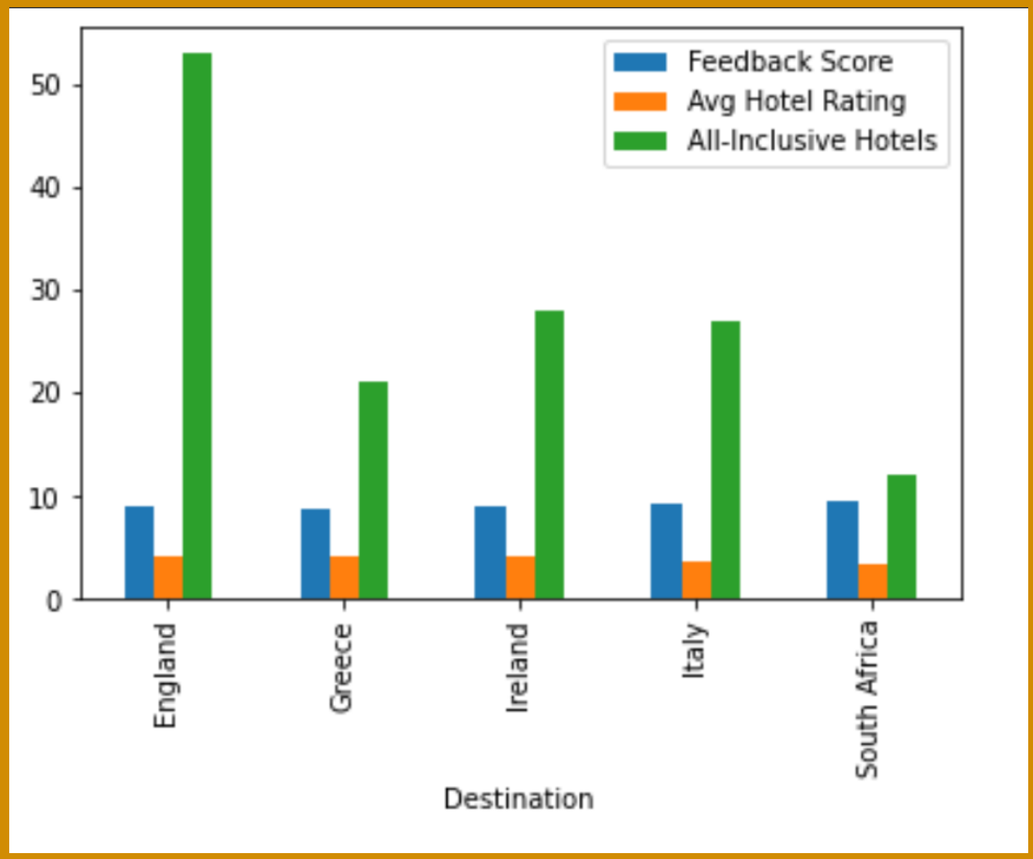
```
Creating useful visualisations is easily done with matplotlib.

import matplotlib.pyplot as plt

9. Is there a correlation between the number of all-inclusive hotels and score?

# Create a scatter chart to show correlation

data.plot.scatter(x="Feedback Score", y="Number of All-Inclusive Hotels")
```



VS Code output showing little to no correlation on the scatter plot

```
Top_5_Destinations = data.nlargest(5,'Feedback Score')
print(Top_5_Destinations)
Top_5_Destinations.groupby('Destination').mean().plot.bar()
```

VS Code script



VS Code output

SUMMARY

This bar graph shows the top 5 holiday destinations with their associated review score