Statistical Analysis

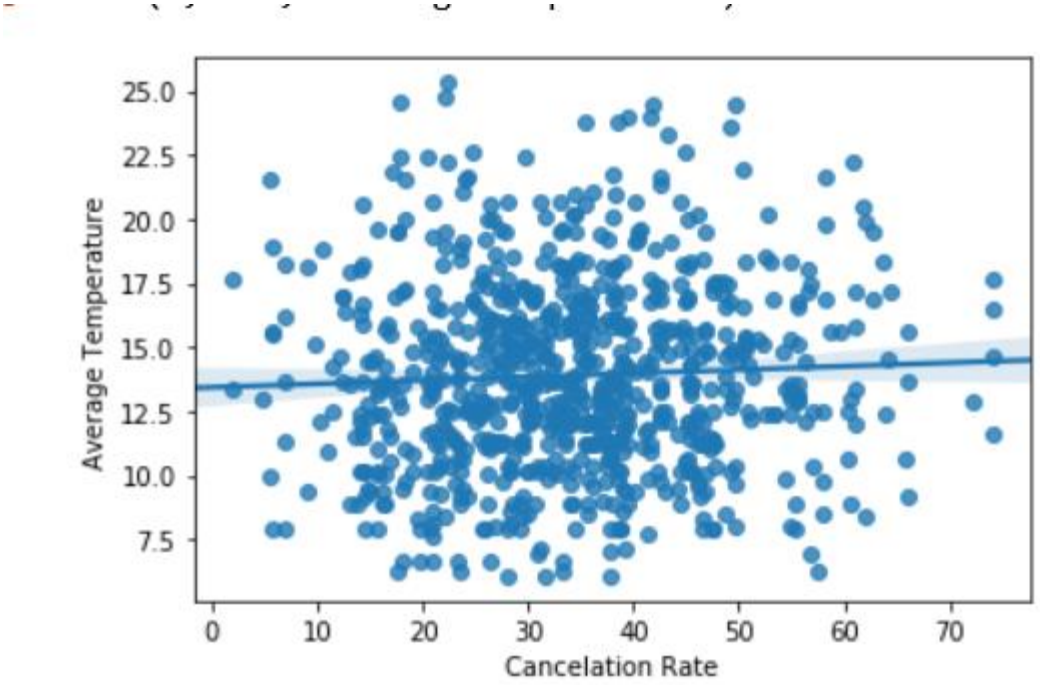
In a data project, statistical analysis is almost the fruit of the data. At this step, we can ask a meaningful question to gain insight or even further investigate clues raised in data storytelling.

With the hotel data I have, there are three questions I want to answer:

1. How does average temperature influence cancelation rate?
2. How does wind speed influence cancelation rate?
3. How does country influence cancelation rate?

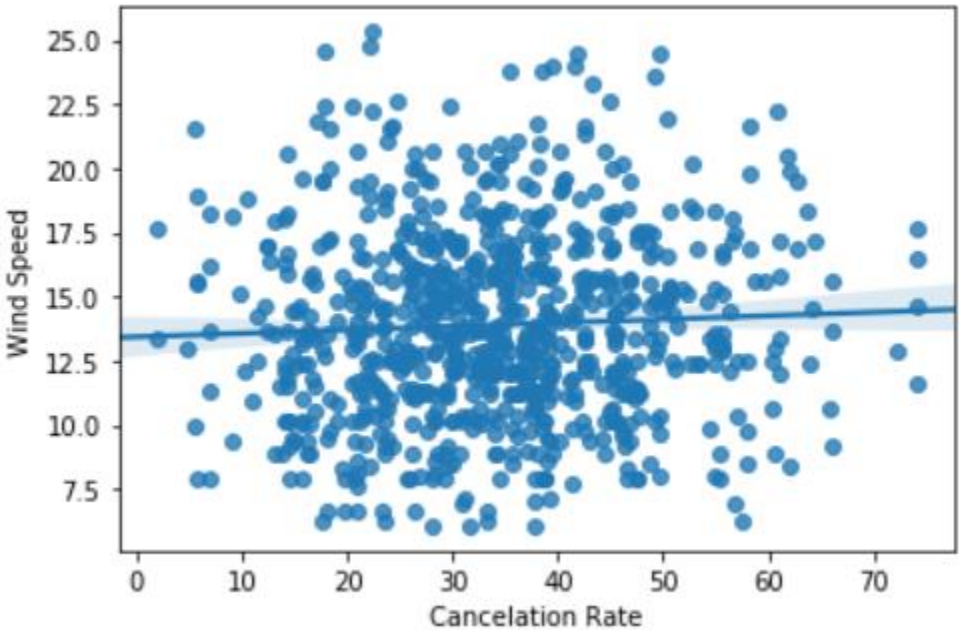
For average temperature vs cancelation rate, the null hypothesis is that the two variables are not correlated, else the Pearson's coefficient will not equal to 0. After obtaining the average temperature and cancelation rate, I used bootstrap inference to loop over 10000 cycles and calculated the Pearson's coefficient and the 95% confidence interval that is equal to [-0.07093038, 0.07070765]. Because the 95% confidence interval contains 0, it is inconclusive whether the two variables have positive, negative, or no correlation. Moreover, as the graph

shown below is the bootstrap points representing cancelation rate vs average temperature:



It is very hard to conclude that there exists a significant correlation between the two variables since the regression line is not very steep.

For wind speed vs cancelation rate, the null hypothesis is that there is no correlation between wind speed and cancelation rate. The alternative hypothesis is that the Pearson's coefficient is not equal to 0. After obtaining the wind speed and cancellation rate after performing bootstrap inference 10000 times, the calculated confidence interval is equal to [-0.06833989 0.06928068]. The result made the test inconclusive because the confidence interval contains both negative and positive numbers. Similar to the average temperature vs cancelation rate, the correlation between wind speed and cancelation rate could be positive, negative, or not correlated at all. In addition, the graph below is very similar to temperature vs cancelation rate:



This graph also illustrates that there is not a significant correlation between wind speed and cancellation rate.

```
list=[]
for i in range(10000):
    sample1=np.random.choice(dom_count['rate'],len(dom_count['rate']))
    sample2=np.random.choice(int_count['rate'],len(int_count['rate']))
    mean1=np.nanmean(sample1)
    mean2=np.nanmean(sample2)
    diff=mean1-mean2
    list.append(diff)
percentile=np.percentile(list,[2.5,97.5])
print('Domestic vs International Cancelation Rate 95% Confidence Interval:'+str(percentile))
```

```
Domestic vs International Cancelation Rate 95% Confidence Interval:[-0.02214565  0.02167502]
```

Lastly, I want to expand on the graph of cancelation rate of each country by examining if the domestic (Portugal) cancelation rate is equal to the international cancelation rate. After extracting the domestic and international cancelation rate, I used bootstrap inference to loop over the data for 10000 times and calculated the difference between domestic and international cancellation rates. The 95% confidence interval is (-0.02134154, 0.02174621) which includes 0.

This made the hypothesis test inconclusive. The graph below also vividly demonstrates that it is not possible to conclude as the two distributions overlap each other heavily.