Mount Rainier Report and Analysis

Prediction of Success Rate for Next Three Months Incl. Weather

February 2021 - Catharina van Veen

# Introduction

Washington state has begun production for their new tourist website. One of the enticing features they would like to add is the statistics of the success rate in which people climb Mount Rainier. Besides just the success rate they have some other questions they need answered which could have an impact on the registration fee, when to advertise to potential tourists, the possibility to hire more park rangers for safety in bad weather conditions and so forth.

They have tasked the National Park Services to collect this data for them to contract out to a Data Science Team.

The client would like a predicted success rate for the next 3 months taking the weather statistics into account. This success rate would be based on the most popularly used route. This will be the scope of this report and analysis.

# Data

For this report I used the data and information about Mount Rainier climbing statistics contained in the provided file, climbing\_statistics.csv, and about the local weather statistics contained in the provided file, Rainier\_Weather.csv.

## Climbing Statistics

The data consists of 4,077 rows and 6 columns all related to Mount Rainier climbing statistics and metrics. Each row represents a party of climbers and provides the following information: date, route, number of failed attempts (incorrectly labeled as attempts), number of successes, and ratio between succeeded attempts and number of failed attempts (incorrectly labeled as success percentage). There are no cells with missing or inconsistent data.

For the purpose of this report, I did not need the ratio between succeeded and failed, so I dropped that column. I added a column to record the total number of attempted (failed + succeeded) and to record the success ratio (succeeded/attempted). I corrected the label of the column for failed attempts. The range of succeeded per party was 0 through 71, the range of failed per party was 1 through 26, and the range of attempted per party was 1 through 76. The success rate ranges from 0 through 93%.

The date was recorded as a string. I converted this to a datetime object. The data covers the period from 4-JAN-2014 through 27-NOV-2015, a period of 693 days. For 391 of these days there exists at least one record. After filtering for the most popular route, the data covers the period from 5-JAN-2014 through 27-NOV-2015, a period of 692 days. For 330 of these days there exists at least one record.

The data has 26 unique routes. Two of these were obvious misspellings of ‘Fuhrer Finger’ and those records were corrected, leaving us with 25 unique route names.

## Weather Statistics

The data consists of 464 rows and 7 columns all related to Mount Rainier weather statistics and metrics. Each row provides the following information: Battery Voltage AVG, Temperature AVG, Relative Humidity AVG, Wind Speed Daily AVG, Wind Direction AVG, and Solar Radiation AVG). There are no cells with missing or inconsistent data.

For the purpose of this report, I did not need the Battery Voltage AVG, so I dropped that column.

The date was recorded as a string. I converted this to a datetime object. The data covers the period from 23-SEP-2014 through 31-DEC-2015, a period of 465 days. There are no duplicate records covering the same day. One day in the series is missing.

The Temperature AVG ranges from 0 through 56.

The Relative Humidity AVG ranges from 10 through 100.

The Wind Speed Daily AVG ranges from 0 through 66.

The Wind Direction AVG ranges from 13 through 325.

The Solar Radiation AVG ranges from 0 through 368.

# Analysis

## Finding Most Popular Route

In figures 1 and 2 it shows that Disappointment Cleaver is by far the most popular route. The number of attempts with route unknown is negligible and will not affect the outcome.

Figure 1



Figure 2



## Analyzing Success Rate Ignoring Seasonal Effects

I calculated the linear regression line, and the results show a r-value of -0.15 and a p-value of 0.006. Since the p-value is well below 5%, this regression is statistically significant. Figure 3 shows success rate per day recorded with the linear regression line. Figure 4 shows that aggregate values for the year 2014 and 2015 show a trend going down. The mean, median, and 1st and 3rd quartile are all lower in 2015.

Figure 3

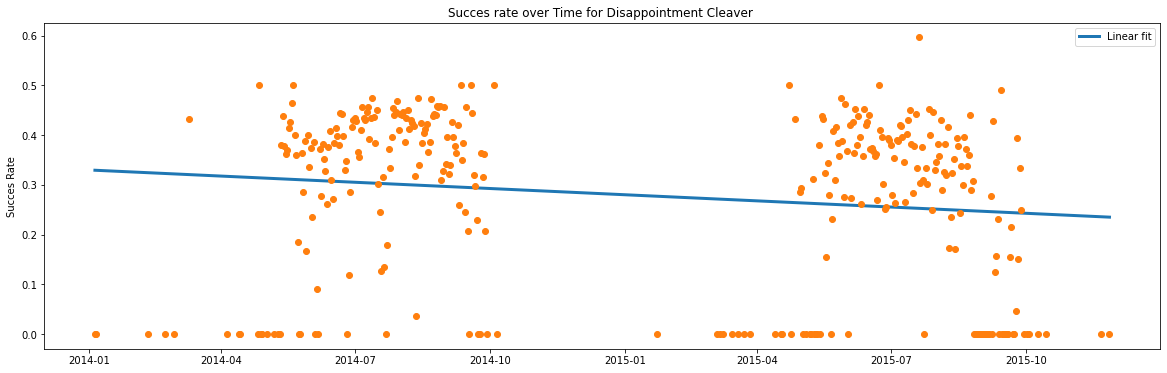
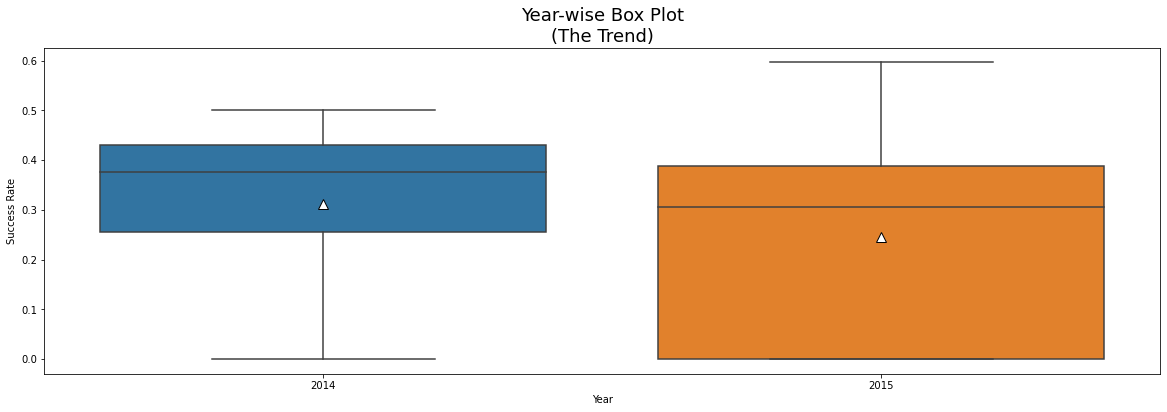


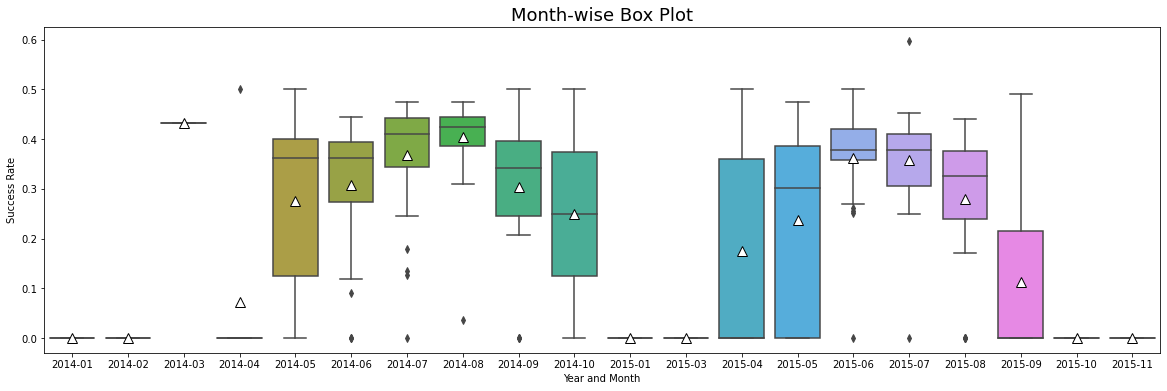
Figure 4



## Analyzing Seasonal Effects

Looking at the box plot of success rate grouped by month we see clear indications of seasonal effect on the success rate (figure 5). Please note that months with no data have been omitted.

Figure 5



## Revisit Trend Not Ignoring Seasonal Effects

Since we now know there are seasonal effects on the data, I revisited the question ‘Is there a trend?’ by applying the Augmented Dickey Fuller (ADF) Test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

### ADF

Results:

ADF Statistic: -5.196724006064108

p-value: 8.928496110239404e-06

Critial Values:

1%, -3.45050711373316

Critial Values:

5%, -2.8704195794076743

Critial Values:

10%, -2.571500856923753

The p-value is obtained is lower than significance level of 0.05 and the ADF statistic is lower than any of the critical values. Therefore, there is reason to reject the null hypothesis. So, the time series is in fact stationary as per the ADF test.

### KPSS

Results:

KPSS Statistic: 0.229586

p-value: 0.100000

Critial Values:

10%, 0.347

Critial Values:

5%, 0.463

Critial Values:

2.5%, 0.574

Critial Values:

1%, 0.739

Based upon the significance level of 0.05 and the p-value of KPSS test, which is higher, there is no evidence for rejecting the null hypothesis in favor of the alternative. Hence, the series is stationary as per the KPSS test.

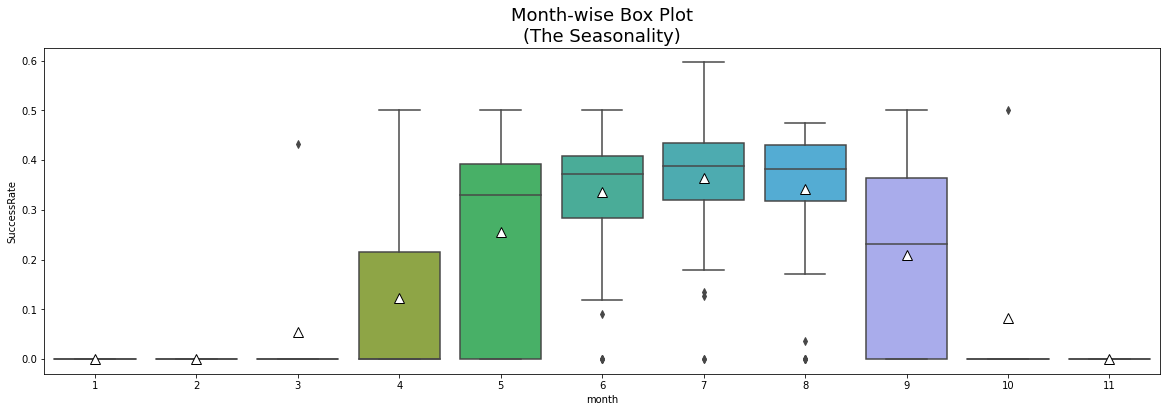
Since both tests conclude that the series is stationary – I concluded that the series is stationary. Stationarity means that the statistical properties of a time series i.e., mean, variance and covariance do not change over time.

## Seasonal Regression Analysis

Since the data is stationary, we can combine the data of the 2 years to determine the seasonal effect.

Fig 6 shows the box plot of the success rate per month for the available years combined.

Figure 6



There were no attempts made in December of either year, so there is no data for December. In November there were only 5 attempts. January, February, and October had 20, 23, and 21 attempts. The other months had significantly higher number of attempts, making the data more reliable. Figure 7 shows the number of attempts per month over both years combined.

Figure 7

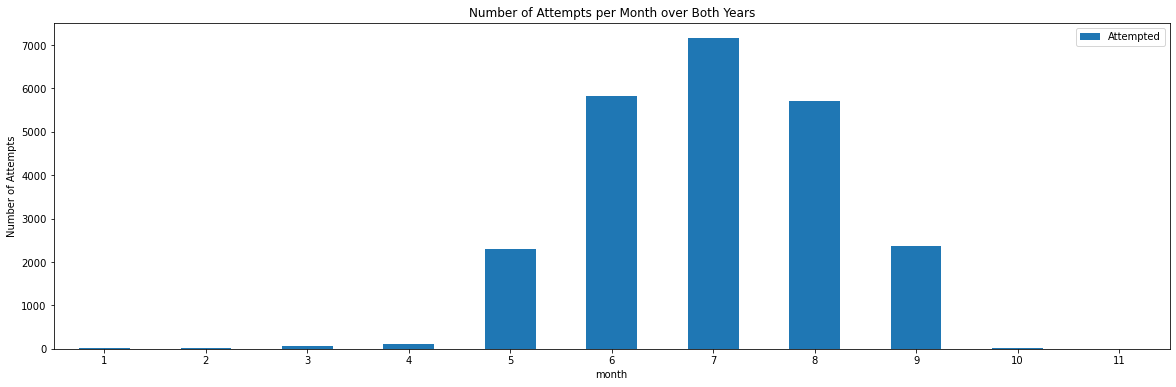


Figure 8 shows the different regression lines, linear, quadratic, and cubic. A visual inspection shows that the cubic regression line is the closest fit. Comparing the mean square errors confirms this. The mean square errors are:

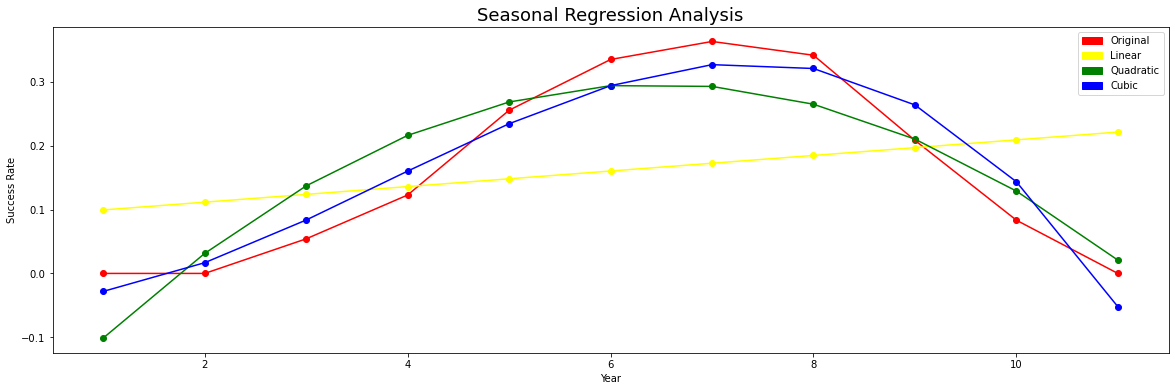
Linear: 0.017745900678414157

Quadratic: 0.00381702029203396

Cubic: 0.001511208731525198

Based on the season, using the cubic regression line, the months of December and January would see a success rate of 0%, February would see a success rate of 2%.

Figure 8

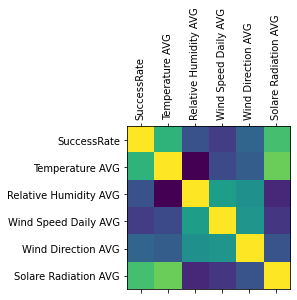


## Including Weather Statistics

I merged the weather data and the climbing data for Disappointment Cleaver. This resulted in records with climbing and weather statistics for 178 days out of 431 days in the range 23-SEP-2014 through 27-11-2015. I dropped the columns 'Failed', 'Succeeded', and 'Attempted', because they are reflected in the success rate.

Figure 9 shows there is a moderate correlation between success rate and temperature AVG and between success rate and Solar Radiation AVG. There is fairly strong correlation between temperature AVG and solar radiation AVG and a fairly strong negative correlation between temperature AVG and relative humidity AVG. There is no significant correlation between the other factors.

Figure 9



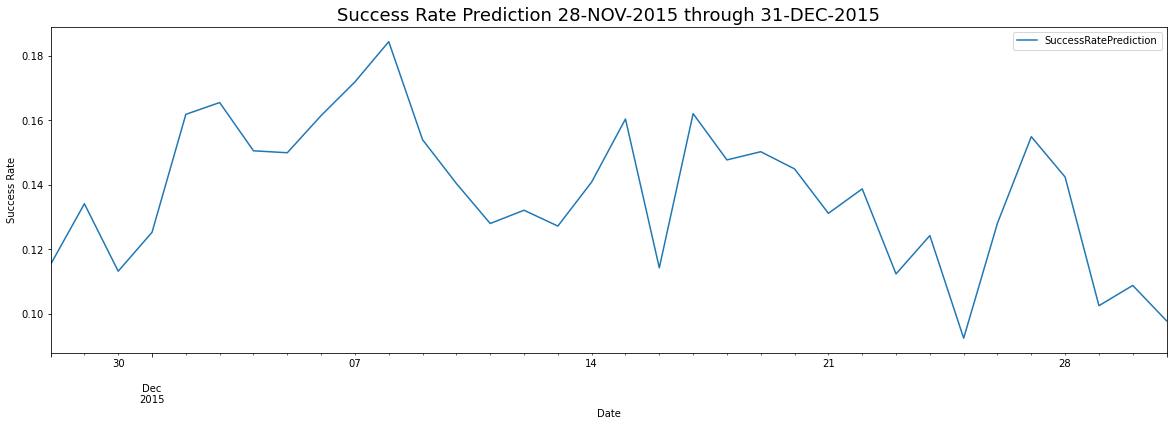
I performed a regression analysis with the factors Temperature AVG, Relative Humidity AVG, Wind Speed Daily AVG, Wind Direction AVG, and Solar Radiation AVG. The r2-score of 0.42771455331742947 indicates there is not a strong correlation. The predicted success rate based on the results of the regression analysis is shown in figure 10. A table with the predicted date can be found below.

# Conclusion

According to the Augmented Dickey Fuller (ADF) Test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test there is no trend over the years. There is clear seasonality.

The predicted success rate based on the results of the regression analysis is shown in figure 10. A table with the predicted date can be found below.

Figure 10



Since no weather data is available for the months of January and February, the prediction is based on the season, using the cubic regression line, January would see a success rate of 0%, February would see a success rate of 2%.

# 

# Predicted Success Rate December

| **Date** | **Success Rate** |
| --- | --- |
| **2015-12-31** | 0.097839 |
| **2015-12-30** | 0.108791 |
| **2015-12-29** | 0.102514 |
| **2015-12-28** | 0.142377 |
| **2015-12-27** | 0.154887 |
| **2015-12-26** | 0.128092 |
| **2015-12-25** | 0.092449 |
| **2015-12-24** | 0.124239 |
| **2015-12-23** | 0.112360 |
| **2015-12-22** | 0.138717 |
| **2015-12-21** | 0.131131 |
| **2015-12-20** | 0.144900 |
| **2015-12-19** | 0.150229 |
| **2015-12-18** | 0.147690 |
| **2015-12-17** | 0.162028 |
| **2015-12-16** | 0.114253 |
| **2015-12-15** | 0.160371 |
| **2015-12-14** | 0.140838 |
| **2015-12-13** | 0.127187 |
| **2015-12-12** | 0.132097 |
| **2015-12-11** | 0.127986 |
| **2015-12-10** | 0.140375 |
| **2015-12-09** | 0.153944 |
| **2015-12-08** | 0.184333 |
| **2015-12-07** | 0.171866 |
| **2015-12-06** | 0.161431 |
| **2015-12-05** | 0.149911 |
| **2015-12-04** | 0.150512 |
| **2015-12-03** | 0.165471 |
| **2015-12-02** | 0.161809 |
| **2015-12-01** | 0.125266 |
| **2015-11-30** | 0.113197 |
| **2015-11-29** | 0.134103 |
| **2015-11-28** | 0.115297 |