**Airbnb Success and Profitability Factors**

Steven Green, Erika Houle and Diana Melendez

**Introduction:**

Airbnb has changed how people can earn money from their properties. It has also affected the hotel industry and how people vacation.

**Inspiration:**

Now that people are returning to vacationing after the impact of the pandemic on the travel industry. We would like to see how has the Airbnb market changed and what projections can we make about the future of this industry.

**Hypotheses:**

Questions:

Growth of Airbnb locations over time and the projections for the subsequent years. (our data ends in 2017).

**Sources of Data:**

Data was collected from Kaggle and Census Data API for Texas. This includes 2 CSVs that included population, poverty…… and Airbnb information such as county, room count, and rate. We also used Google Maps API for creating heatmaps and map markers for our data.

**Data Cleaning and Exploration:**

**Analysis:**

Questions addressed:

**ANOVA: Economy (GDP per Capita):**

After investigating the ANOVA for overall Happiness Score, we ran it with all of the different variables that made up the Happiness Score. In table 1 we can see that all of the variables were statistically significant. It is important to note that for the variable Economy, the statistically significant result was fragile. By looking at the series of means in figure 4, the year 2015 is a very slight outlier compared to the rest of the years. Despite how inconsequential this

may appear, we found that after removing this year from the test, there was no longer overall statistical significance for Economy.

**Variable**

Happiness Score  
Economy (GDP per Capita) Social Support  
Health Life Expectancy Freedom  
Trust (Government Corruption) Generosity

**P-Value**

0.9522 0.0163 4.16E-07 1.95E-12 3.40E-13 2.37E-30 1.39E-08

page4image62850176page4image62856704page4image62855744page4image62863040

**ANOVA: Health Life Expectancy:**

Through ANOVA testing we found that another significant factor of overall happiness was the Health Life Expectancy variable. We also tried manipulating the test by removing the year 2017 in figure 5, as it was the outlier for this variable, to see if it would cause the same result as the Economy test. Once it was removed, the data was still statistically significant.

**Top & Bottom Countries:**

After investigating the difference between different variables across the years we decided to dive a little deeper and look  
into the major difference between  
countries that were repeatedly in the  
top and bottom 10% based on their  
Happiness Score. We found three total  
countries that were in the top 10%  
every year and three countries that  
were in the bottom 10% every year.  
The top three countries were Denmark,  
Norway and Iceland. The bottom three  
countries were Rwanda, Afghanistan,  
and Tanzania. We then graphed each  
variable to see how each of these six

countries changed over the years and to determine if there was any pattern we could visualize. As you can see in figure 6 there is a visual difference between the Happiness Score for the top three countries and bottom three countries. It was difficult to visualize if there was any consistent variation across all 6 of the countries when they were graphed together, so we also graphed them separately in figure 7 and figure 8. In figure 7 it is easy to see that each top country certainly had variation over the years, but they all did not increase or decrease at the same time. Similarly, in figure 8, the bottom countries also varied across the years.

We also were able to compare each of the individual variables that were summed into the Happiness Score and how they varied based on the top and bottom countries. As you can see

in figure 9 it is very apparent that all three of the top countries' Health and Life Expectancy dropped from 2019 to 2020 whereas the bottom three countries have risen from 2019 to 2020. It is also worth noting that these values are quickly approaching each other, yet the bottom three countries remained in the bottom 10% and the top three countries remained in the top 10% based on Happiness Rank. Similarly figure 10 shows how

these six countries fared in terms of economy over the six years included in the study. Again, as the years go on, the top three countries are slowly decreasing whereas the bottom three continue to rise. This analysis of how the top and bottom countries changed over time lead us to dig deeper into our data to see what variable correlated most with Happiness Score.

**T-tests:**

We decided to then perform t-tests on the overall Happiness Score, and the two significant variables, Health and Economy. We did this by comparing the top 10 countries Happiness Score of 2015 to 2020, and then ran the same test for the bottom 10 countries. When looking at Happiness Score, both tests produced p-values greater than .05. Therefore the changes over time are not statistically significant, and we rejected the alternative hypothesis, accepting the null.

When running the test for Economy and Health, these tests all produced p-values less than .05. Therefore the changes were statistically significant, rejecting the null hypothesis and accepting the alternative. The result of each test can be seen below, in table 2.

**P-Values**

**Top 10 Countries Bottom 10 Countries**

**Correlations:**

**Happiness Score**

0.6576 0.6796

**Economy (GDP per Capita)**

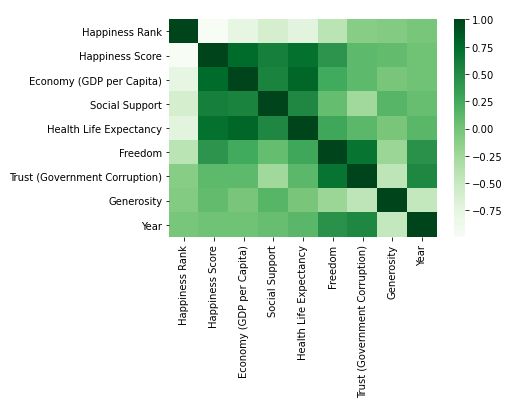
1.73E-8 5.43E-7

**Health Life Expectancy**

5.94E-10 0.0005

Next we wanted to look at correlation of Happiness Score to each individual variable. As we can see here in figure 11, Economy and Health Life Expectancy have the strongest correlations.

From there we decided to look into these variables further to see how strong the correlations are.We started off by creating scatter plots for each variable to Happiness Score, and then ran linregress. For both of these figures we can now visually see the positive

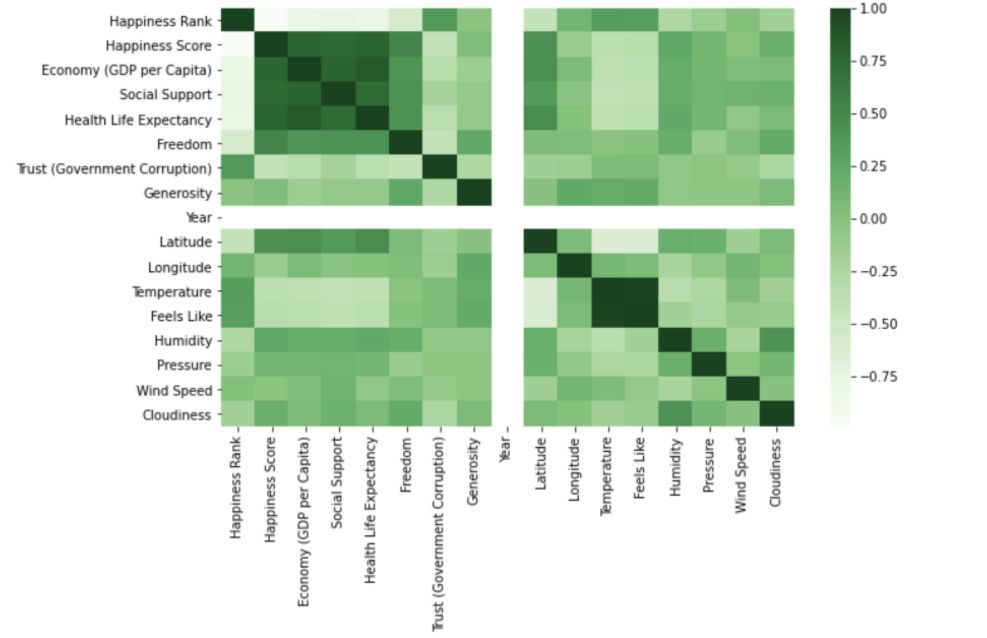


correlations they have. Next, to see just how strong the correlations were, we ran stats models. For figure 12 we got an r-value of .575, and for figure 13, we got an r-value of .510. Since both r-values were above .5, they have a strong positive correlation. We also ran the stats models on the other variables. All of them had a positive r-value that was below .5 and therefore have a weak correlation with Happiness Score as can be seen by the results in the Appendix.

**Correlations with Weather:**

As mentioned above we wanted to cross examine our data with an outside data source. We combined the weather data for each country in the 2020 World Happiness Report Data set and compared it with each variable that was included in the 2020 World Happiness Report. As you can see in figure 14 there are not very many strong correlations between any of the weather variables and the World Happiness Report variables. When looking at the heat map you can see a slight positive correlation between Latitude and Happiness Score and a slight negative correlation between Temperature and Happiness Score.

In order to investigate these correlations deeper, we were able to create a scatter plot and linregress models for each weather variable with their Happiness Score. After reviewing each R-value it became apparent none of the weather variables had a strong correlation with Happiness Score. We decided to include the two variables with the strongest correlations. As you can see in figure 15 there is a weak relationship between Happiness Score and Temperature. The calculated r-values for the simple linregress is ​0.44 which shows that there is a weak positive correlation between Happiness Score and Latitude.



Similarly, the relationship between Happiness Score vs Latitude as you can see in figure 16 shows a slight negative correlation. ​The calculated r-values for the simple linregress is ​-0.35 which shows that there is a weak negative correlation between Happiness Score and Temperature.

**Prediction Model:**

For our Prediction Model we started off by running the stats models on all of the variables from the World Happiness Report versus Happiness Score. This gave us an Adjusted R-squared of .714, as seen in table 5. As mentioned before, the Happiness Score is a sum of all of the variables, so it was expected that the r-squared would be high. We also looked at the p-values for each variable, and all were at zero, or very close to it. From this information we gather that none of the variables are having a negative effect on correlation.

From there we created two scatter plots comparing the Predicted versus Actual Happiness Score, and the Predicted versus Residual Happiness Score. In figure 17, we can see that the lower Happiness Scores were over predicted because most of the plots are above the regression line, while the higher Happiness Scores were under predicted because all of the plots are below the regression line. The predictive modeling works best on a Happiness Score that falls between four and seven. These plots are pretty evenly distributed above and below the regression line, or on it. Figure 18, shows that Happiness Scores between four and seven have the most points that are closest or on zero. This further implies that our predictive modeling works well for this Happiness Score range.

**Prediction Model with Weather:**

Using our correlation heat map we used a multivariate regression prediction model to investigate if multiple weather variables  
could predict Happiness Score. We  
determined that the “best” model  
included Latitude, Humidity, and  
Cloudiness. When we attempted to  
include any other weather variables the  
Adjusted R-squared value actually  
decreased. In table 6 it is important to  
point out the Adjusted R-squared value  
of 0.210 and different p-values for each  
variable included.

Overall, we determined that our prediction model including the weather variables would only be accurate about 21% of the time, which indicated that our regression model is not the best. In figure 20 it is easy to see that our predictions do not cluster around the

actual values (pink line) and there is a large discrepancy between the two values. This is also visualized in figure 21 where the residual values do not cluster around the 0 value, further proving that our regression model cannot be validated.

**Conclusion:**

For our hypotheses we can now conclude that we were correct that happiness has increased over time, weather does not positively affect happiness, and that our ability to predict happiness is questionable because it is dependent upon the variables.

**Limitations:**

There were a few limitations that we came across in our exploration of this dataset. These were in part due to our own lack of mastery over the method, but also due to the data itself. The dataset was limited because it was trying to quantify a feeling that everyone defines differently. While it may at first seem like the responses to the questions given were prone to subjectivity, we instead believe that the questions asked were. The dataset was also not uniform across all 6 years. Some countries were only reported for one year, and we didn’t use the raw dataset for this reason.

**Future Work:**

In the future we would like to expand our weather data by accumulating it over the same 2015-2020 time period and see if there really is any correlation between weather and happiness.. We would also bring in other factors such as education, housing, etc. to see if there is any connection between them and Happiness Scores.

**References**​:

* ●  **Census API**
* ●  **Google Maps API:**​ https://cloud.google.com/maps-platform/
* ●  **Inspiration:**
* ●  **Kaggle Notebooks:**