2019

PROJECT 1

GROUP 6

MELISSA MORGAN

GODWIN THOMAS

JULIA THOMPSON

BLAKE SKINNER

Group 6

HAPPINESS & SUNSHINE

"When I was 5 years old, my mother always told me that happiness was the key to life. When I went to school, they asked me what I wanted to be when I grew up. I wrote down 'happy'. They told me I didn't understand the assignment, and I told them they didn't understand life."

John Lennon

DO THE HAPPIEST PEOPLE REALLY LIVE IN THE SUNNIEST CITIES?

Is there a direct correlation between a city's happiness score and the amount of sunshine a city experiences each year?

PROJECT HYPOTHESIS

Null Hypothesis:

The amount of cloud coverage, and reversely the amount of sunshine, has no correlation to the happiness score of a city.

Alternative Hypothesis

The amount of cloud coverage, and reversely the amount of sunshine, does affect the happiness score of a city.

SOURCES

- Google Geocoding API
- Dark Sky API
- Happiest Cities in America
  Mar 11, 2019 | Adam McCann, Financial Writer

Description of the data

Happiest Cities in America Study

The study resulted in a list of 182 cities ranked by each city's total happiness score.

Study Methodology: WalletHub compared 182 of the largest cities — including the 150 most populated U.S. cities, plus at least two of the most populated cities in each state — across three key dimensions: 1) Emotional & Physical Well-Being, 2) Income & Employment and 3) Community & Environment.

These categories were evaluated using 31 relevant metrics, each graded on a 100-point scale, with a score of 100 representing maximum happiness. Each city's weighted average across all metrics was used to calculate its overall score and used the resulting scores to rank-order the sample.

Google Geocoding API

Per the documentation, the Geocoding API is a service that provides opportunity for geocoding--"the process of converting addresses (like a street address) into geographic coordinates (like latitude and longitude), which you can use to place markers on a map, or position the map."[1]

Dark Sky API - Time Machine Request

Per the documentation, a Time Machine Request for the Dark Sky API "returns the observed (in the past) or forecasted (in the future) hour-by-hour weather and daily weather conditions for a particular date".[2] For purposes of this project, we pulled in historical daily weather condition information for our test cities.

METHODOLOGY

Gathering and Cleaning the Data

The Happiest Cities in America study provided all of the data needed for the "happiness" portion of our study – city and state information with corresponding happiness scores and rankings.

As for the weather portion of our study, we researched various online sources and APIs.

The biggest factor in this process was looking into the specific variables/types of data available from each source and whether or not that source could provide historical data. We decided to use the Dark Sky API as it contains past and present daily weather information (as opposed to other APIs that only provide current and hourly data) and it has sufficient cloud coverage and UV Index information.

In order to pull the weather data from Dark Sky's API, we needed the latitude and longitude info for each city. We utilized the **Google Geocoding API** to acquire this data for all 182 cities published in the Happiness study -- iterating through our dataframe and adding the respective lat/long information to the same dataframe.

We then used the lat/lng information to hit the **Dark Sky API**. We pulled the daily weather information the following six variables:

- **Summary**: A human-readable text summary of this data point.
- **Icon**: A machine-readable text summary of this data point, suitable for selecting an icon for display.
- **Sunrise time**: The UNIX time of when the sun will rise during a given day.
- **Sunset time**: The UNIX time of when the sun will set during a given day.
- **Cloud cover**: The percentage of sky occluded by clouds, between 0 and 1, inclusive.
- **UV Index**: The UV Index.

[1] https://developers.google.com/maps/documentation/geocoding/start

[2] https://darksky.net/dev/docs

We selected these six variables was based on our hypothesis – that cities with more sunlight have higher happiness scores. Therefore, we extracted the variables surrounding sunlight to test our hypothesis. We successfully this data for each of our 182 cities, beginning on January 1, 2018 and ending on December 31, 2018.

After gathering our data, we addressed the following issues:

Geographic Location: Continental United States vs. All 50 States

- We removed Hawaii (due to its geographical location so far West) and Alaska (due to its extremes in daylight hours) from the dataset to narrow our focus to the continental U.S. only.
- The discarded data included two cities in Hawaii and two cities in Alaska.
- Once removed, we did not re-rank the cities.

Missing Data from Dark Sky API

- When pulling in the daily data from the Dark Sky API, there were 15 days that did not include cloud coverage or UV Index information. We decided to delete these rows entirely from the dataset.
- Days deleted: Las Cruces, NM (2) and Casper, WY (13).
- These days represented less 0.00023% of the collected data.

Analyzing the Data

For this project, we focused our data analysis on the following techniques:

Visualizations

- Scatter plots
- Heatmaps (Geo & SNS)
- Bar graphs

Statistical Analysis

- T-tests
- Regression analysis
- Correlation

DATA LIMITATIONS

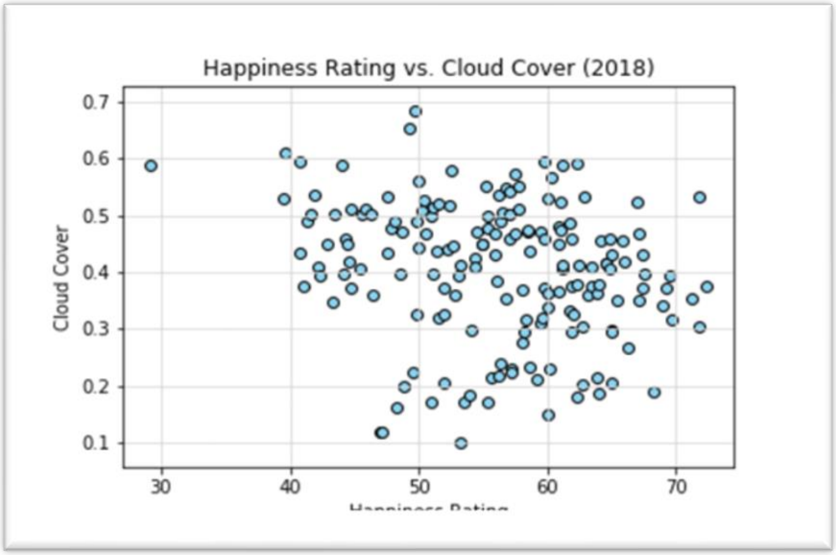
- Happiness, while many factors were looked at in a quality attempt to define overall happiness, it is still subjective. Joy and happiness are defined differently for everyone. The Happiness study is the Author's opinion based off data collected from many reputable sources.
- The Happiness study's data only looks at larger cities and doesn't consider smaller cities. This fact alone could completely skew the data.
- Also, the study did not consider metropolitan areas as a whole – which may have resulted in a sample that provided significantly more conclusive data for the entire population. For example,

the quality of life factor could be skewed by suburban life for certain cities. Ideas like collective average IQ and political management of different cities could affect the happiness of the population as well.

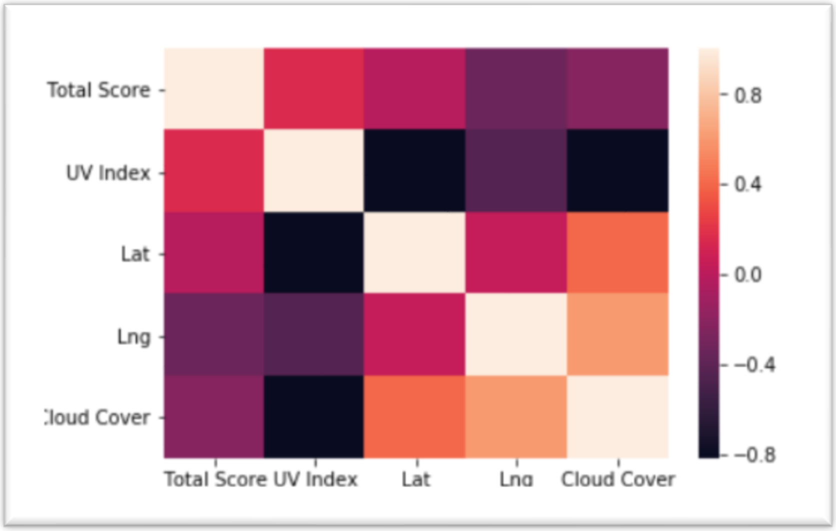
- We did not perform any analysis on the weather data itself. We collected information for an entire calendar year, but we have no way of knowing whether the weather data obtained is truly representative of the typical year in each city. Therefore,
- Potential data limitation that we recognized and quickly found a solution: Our dataset contained six cities with the same name in two states (i.e. Columbia, SC and Columbia, MD). This was important to note when performing the groupby function and consolidations on the data. To avoid combining the data for two separate cities, we grouped by Overall Rank or Total Score rather than city name.

DATA ANALYSIS

Cloud Cover and Longitude Analysis



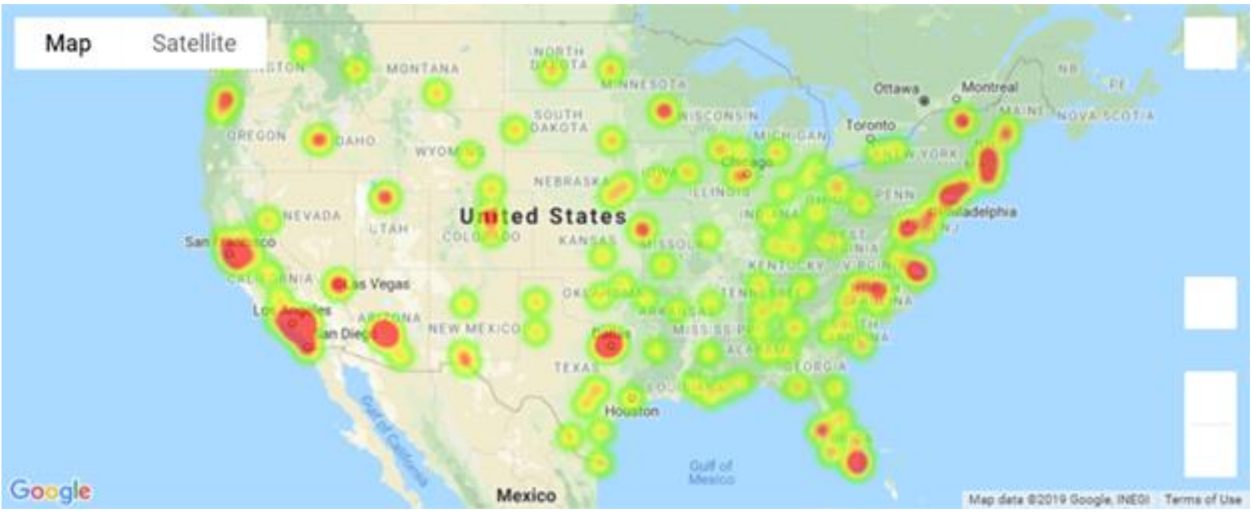
Based on this scatter plot visualization, there appears to be no correlation between the "Happiness Rating and Cloud Cover". We employed a series of heatmaps to further investigate the correlation between longitude, cloud cover, and overall happiness rating.



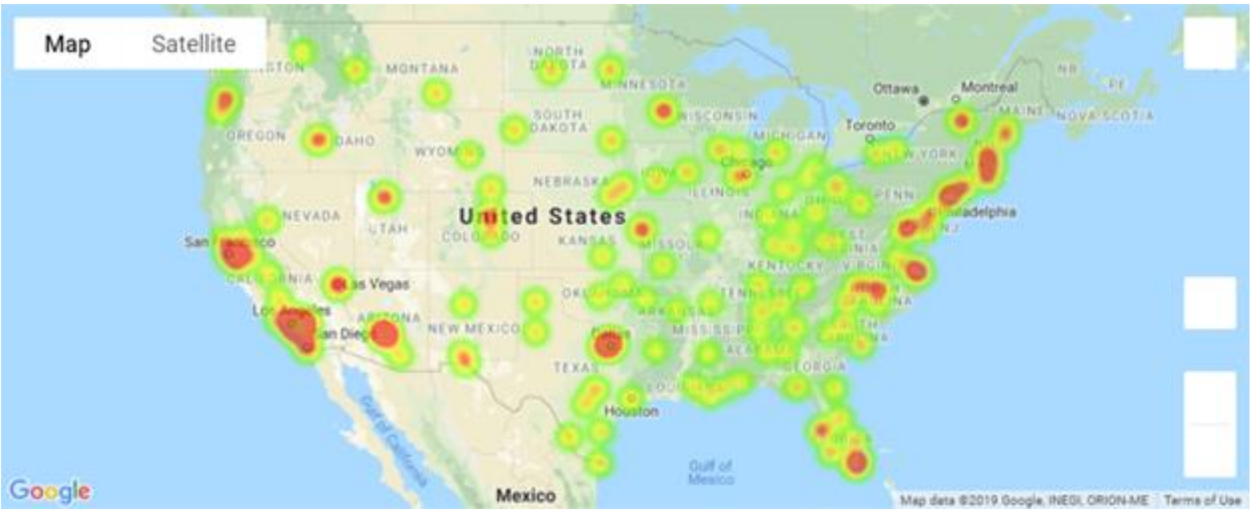
This SNS heatmap illustrates that the strongest correlation exists between Total Score/Happiness Rating and Longitude. According to our regression results, Longitude explains 11.2% of the variation in the happiness rating for cities across the continental United States. The second strongest correlation is Happiness Rating and Cloud Cover. After statistical analysis, cloud cover explains 5.3% of the variation in the happiness rating.

We then decided to use maps to get a visual representation of the location of our 178 cities. The following heatmaps display the location of the happiest cities in the continental U.S. based on happiness rating, location, and subsequently those cities which experience cloud cover.

Happiness Rating and Location:



Happiness Rating, Location, and Cloud Cover:

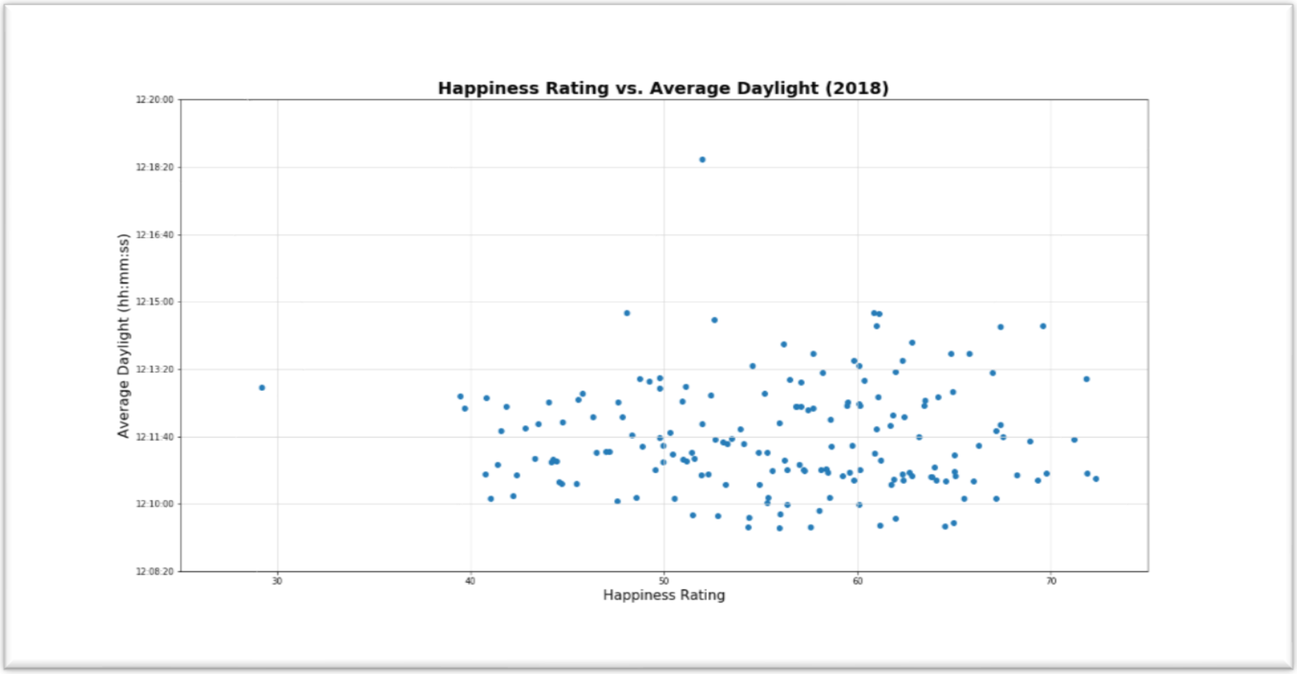


Findings: As shown in the SNS heatmap, Happiness and Longitude have a negative correlation which explains why there appears to be a larger concentration of "Happier Cities" along the east coast of the United States. The addition of cloud cover has a minimal effect on happiness.

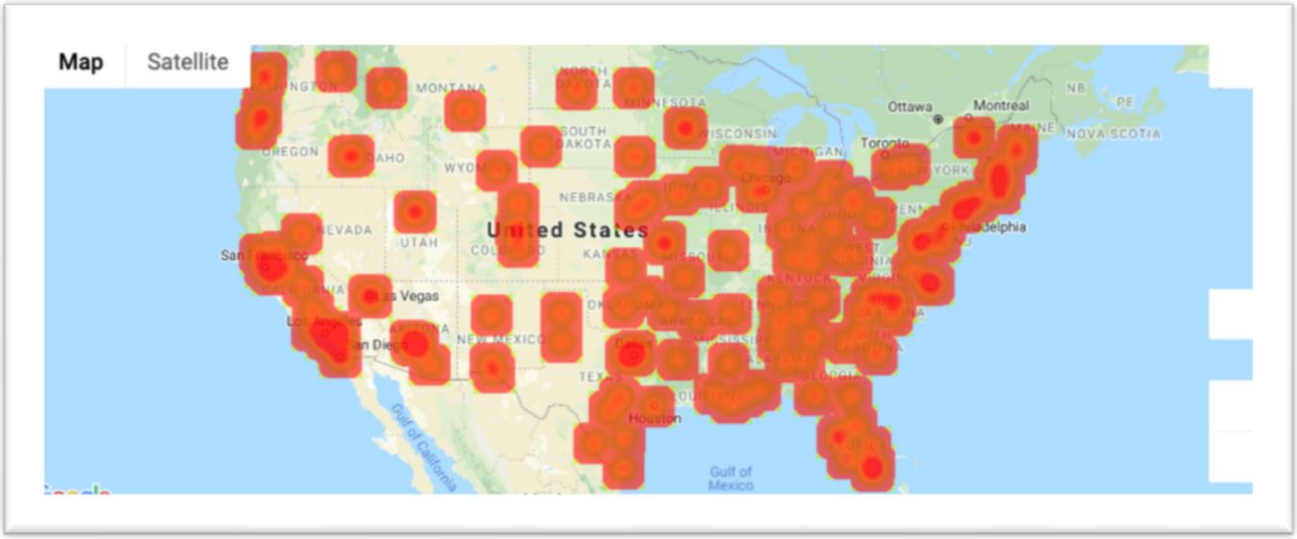
Hours of Daylight

- To find average sunlight hours, the dataframe was first check to find any NAN objects and a drop na was executed on the code to make sure all columns we filled.
- Next, lists were initialized to store the sunrise and sunset unix code, after which, the unix time code was converted to a date time format and stored in a separate sunrise and sunset list. Next, the sunset was subtracted from the sunrise to find the total daylight hours per day. Iterrows was used to populate a new column onto the dataframe with the hours, minutes, and seconds.
- Next, total seconds per day was calculated using a list comprehension, and a new column on DataFrame was created with total seconds.
- Groupby was used to group total score and find average daily light seconds per city.
- Another list comprehension was used to find average time in hours, minutes, and seconds.
- Finally, a scatter plot and heatmap was displayed, and the R2 was calculated to analyze the correlation.

Scatter Plot – Average Daylight Hours



Heat Map – Average Daylight Seconds

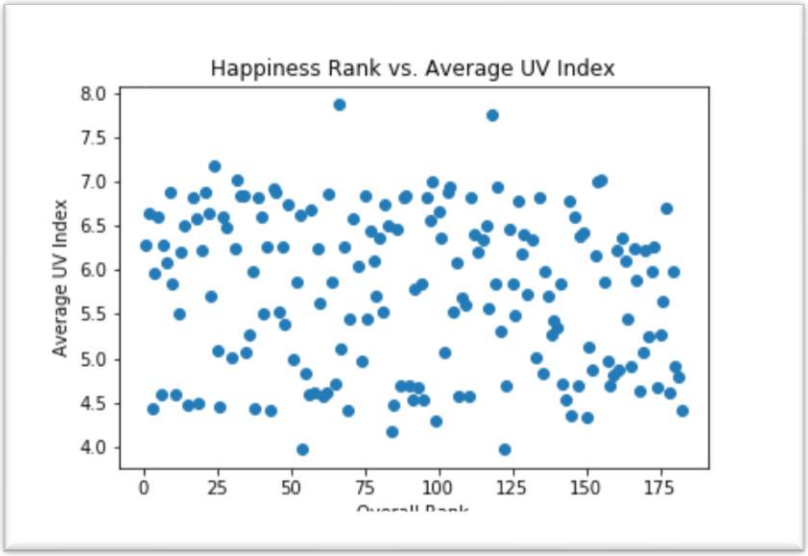


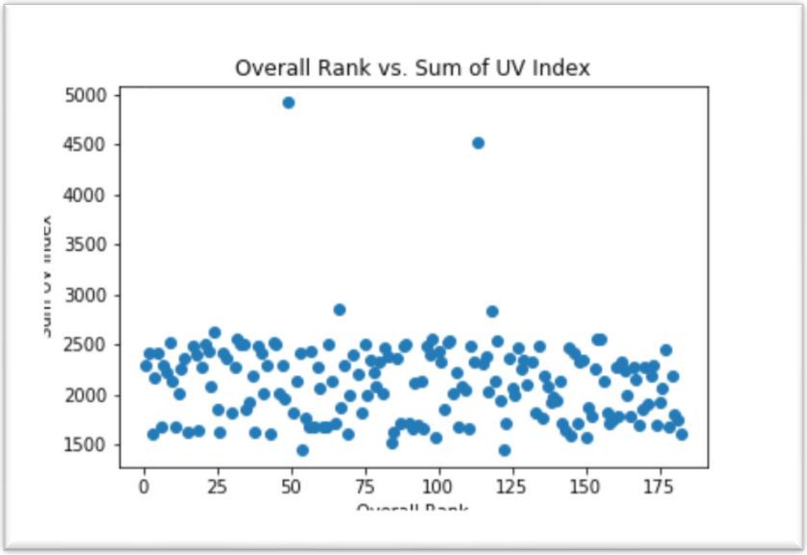
Results/Conclusions

- There was no correlation between daylight hours and the happiness Index
- Apart from some outliers, most cities ranged between 12 and 12 hours and 15 minutes of daylight.
- The city that had the highest average daylight hours was in the middle of the other cities

Additional Scatter Plots

We knew from the SNS Heatmap that UV Index probably would not have a strong correlation to a city's overall happiness rank. However, to be sure, we compared the overall happiness rank (1=Best) against the UV Index (average and sum) per year. The scatter plots below illustrate the results.





Findings: As anticipated, there is not enough evidence to conclude that the average UV Index has a significant impact on a city's overall happiness rank.

CHANGE IN APPROACH TO THE DATA ANALYSIS

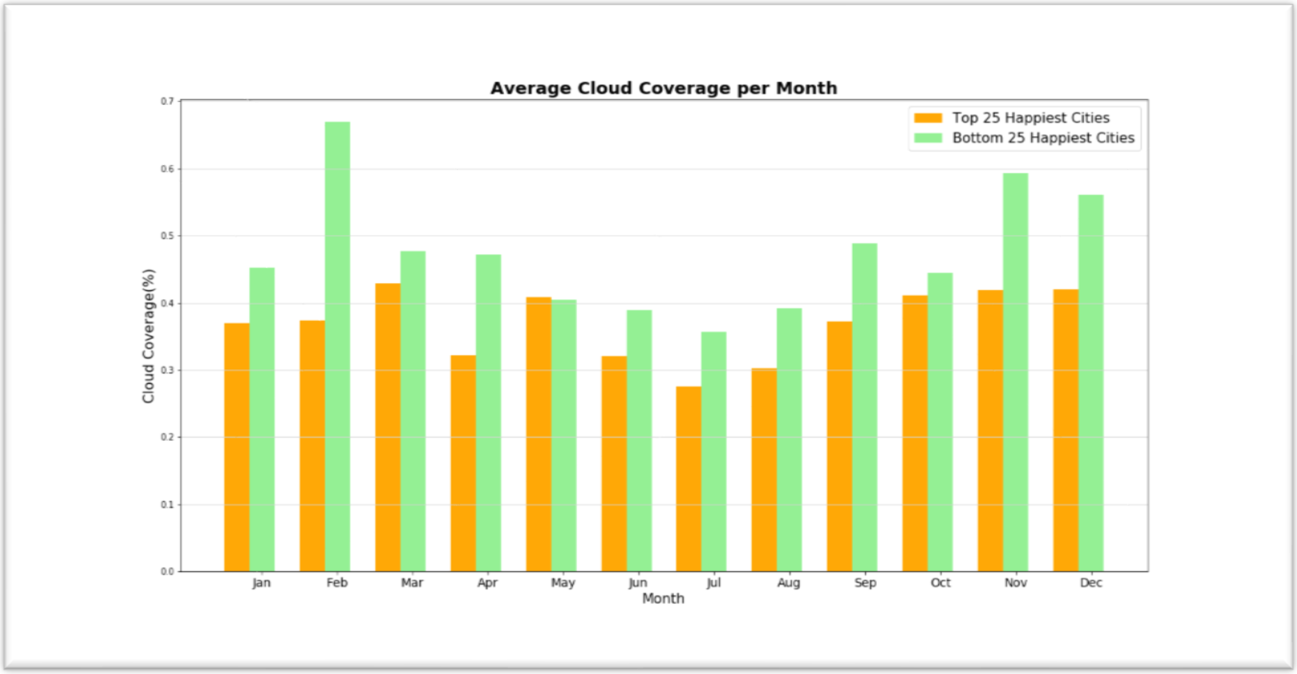
Top 25 & Bottom Happiest Cities (based on happiness rank)

We shifted our focus – perform analysis on the top 25 and bottom 25 cities based on happiness rankings.

Methodology

- Created two separate dataframes – one containing the top 25 happiest cities and one containing the bottom 25 happiest cities
- Grouped each dataframe by month.
- Ran comparisons for the two datasets on cloud coverage and number of clear days.
- Created visualizations for the results.
- Performed two sample t-test on each.

Cloud Coverage



Null hypothesis: The average monthly cloud coverage for the top 25 happiest cities is equal to (or not different from) the average monthly cloud coverage for the bottom 25 happiest cities.

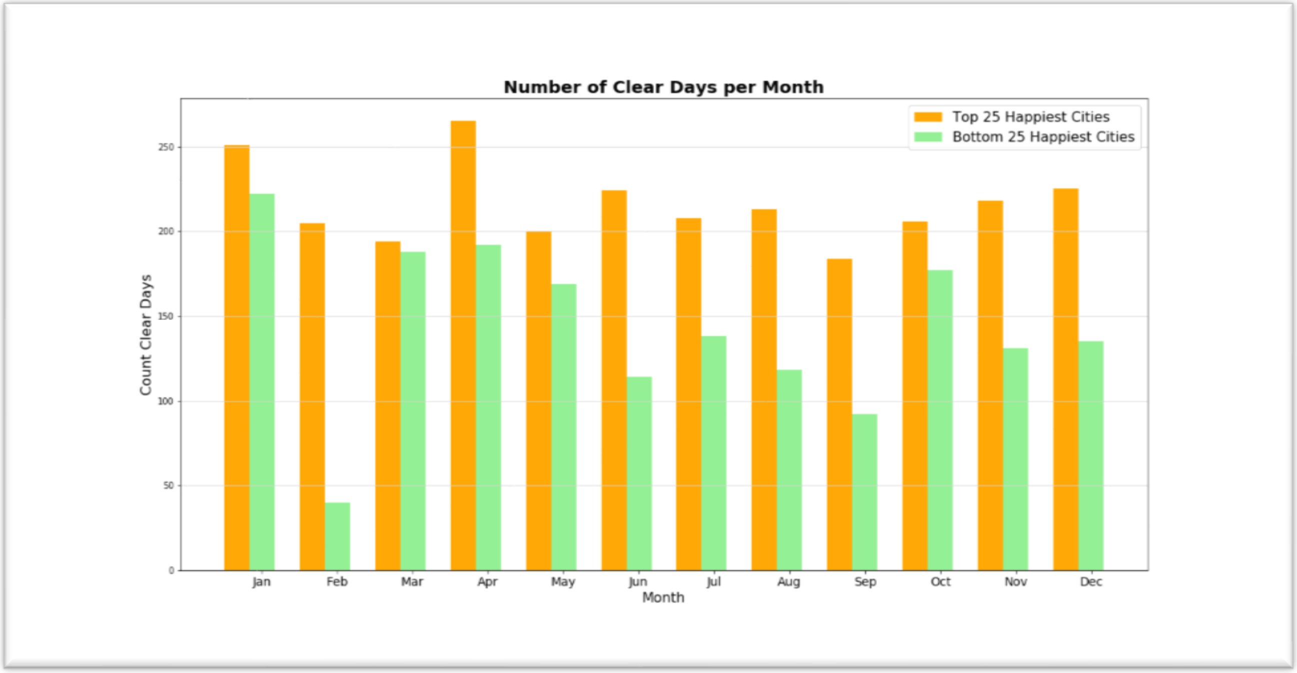
Alternative hypothesis: The average monthly cloud coverage for the top 25 happiest cities is different than the average monthly cloud coverage for the bottom 25 happiest cities.

Statistical Analysis (Two-Sample t-Test)

Results Two-Sample t-Test	Statistic/ T-Score	P-Value
Average Cloud Coverage per Month	3.483305	0.002777

Conclusion: With a p-value is .003, which is less than alpha of .05, we reject the null hypothesis and accept the alternate hypothesis that the average monthly cloud coverage is different for the top 25 cities in happiness and the bottom 25 cities.

Number of Clear Days



Null hypothesis: The average number of clear days per month for the top 25 happiest cities is no different than the average number of clear days per month for the bottom 25 happiest cities.

Alternative hypothesis: The average number of clear days per month for the top 25 happiest cities is different from the average number of clear days per month for the bottom 25 happiest cities.

Statistical Analysis (Two-Sample t-Test)

Results Two-Sample t-Test	Statistic/ T-Score	P-Value
Number of Clear Days per Month	4.605415	0.000317

Conclusion: The number of clear days per month for the top 25 happiest cities is statistically significantly different than the number of clear days per month for the bottom 25 happiest cities. Looking at the numbers, we can confidently claim that the number of clear days per month for the top 25 happiest cities is greater than the number of clear days for the bottom 25 happiest cities

PROJECT CONCLUSION

After statistical analysis of the overall dataset, regression and significance testing have not provided a strong enough correlation so therefore we cannot reject the null hypothesis that states:

The amount of cloud coverage, and reversely the amount of sunshine, has no correlation to the happiness score of a city.

However, through a narrowed scope of investigating the top 25 happiest and bottom 25 least happiest cities their amount of cloud cover differs greatly. Through statistical evidence in this narrowed scope we can suggest the alternative, which is that sunshine and cloud cover can affect happiness ratings for a given city. With significance testing we were able to find a p-value of 0.003 that supports the major difference in cloud cover between the top 25 and bottom 25 cities.

"The purpose of our lives is to be happy." Dalai Lama