# **Hawaii Airbnb Listing Analysis**

# **Sha Liu**

# **4-23-2021**

# **Introduction**

In this report, I have analyzed the Airbnb listings dataset for Hawaii from 2008 to 2021 on Google Cloud Platform and learned about the data pipeline and lifecycle. I was able to use AI Platform Notebook instance on Google Cloud to run my analysis. I was able to perform each step are list in the USGS lifecycle model [1], which includes Plan, Acquire, Process, Preserve and Publish/Share Airbnb listing analysis.

Graphical user interface, application

Description automatically generated

# **Background**

I really like the module talks about data lifecycle and pipelines and would love to strength my skills I learned from class; I think it will benefit my career as I am pursing a data engineer role at the moment. At the same time, I want to get hands-on experience on cloud computing, so I chose to use Google Cloud as my platform. I chose to work on Hawaii Airbnb Listing analysis because my favorite vacation place is Hawaii and I have been to Maui before and love to explore other islands in Hawaii as well. I am interested in getting an Airbnb in Hawaii next time I visit. Therefore, I would like to understand what Airbnb listings are out there and what are the statistics I can pull out it. It will definitely help me make well-informed decisions when I choose my future Airbnb in Hawaii. Luckily, I was able to find the public dataset on Inside Airbnb website and perform data analysis.

# **Methodology**

For Google Cloud, I was able to get free $300 credit for first time user. I started by setting up the environment for Jupyter Notebook on Google Cloud. I was able to use Jupyter Notebook AI to run my notebooks. I know that I can utilized what I have learned when I completed "Visualize the 10,000 Bitcoin Pizza Transaction Using BigQuery and AI Notebooks (Links to an external site.)" qwiklab on one of the class modules.

Firstly, I created a project named “Airbnb Reviews”, shown below:

Graphical user interface, application

Description automatically generated

I then created a Notebook Instance and selected Python 3, shown below:

Graphical user interface, text, application, email

Description automatically generated

I opened the JUPYTERLAB and installed necessary packages were needed to run my notebook in the google could terminal shown below:

Text

Description automatically generated

Then I created a notebook named “airbnb\_listings\_reviews” and started the data lifecycle. I have attached the whole notebook in the end.

Graphical user interface, text, application, email

Description automatically generated

# **Results**

I was able to load the Hawaii Airbnb Listing dataset successfully and performed the following steps.

* load data
* data summary
* data cleansing
* exploratory data analysis
* data visualization

1. **Load data**

The data format on the Inside Airbnb website is in csv.gz. I loaded the dataset by using wget the url and read it csv file into pandas dataframe.

1. **Data summary**

The dataset contains 74 columns and 22,031 listing records. I wasn’t able to find the meta data, but the columns names and data types are pretty straight forward. I performed df.info() function to display all the columns and their data type as well. The dataset contains listings started from 7/21/2008 to 2/6/2021.

I also checked the null values and show the top 20 columns have missing values to better understand the data. Luckily, the missing values columns do not impact my analysis. The columns I decided to analyze are the following and they do not have missing values:

* **neighbourhood\_group\_cleansed**: it contains 4 groups showing below: Kauai, Honolulu, Maui, and Hawaii. The Hawaii groups contains all small islands around Hawaii.
* **price:** I would like the to the price among all neighborhoods
* **room\_type:** I like to know which room type are the most popular
* **property \_type:** I like to know which property type are the most popular
* **review\_score\_rating**: I would like to the how are the ratings distributed
* **name**: I would like to know what are the most popular words in the listings

1. **Data cleansing**

I wanted to analysis how the price varies among neighhoods, so I checked the data types of price, and it is in string. I converted the price from string to float, so it is easy for perform statistical analysis.

1. **Exploratory analysis and Visualization**

The distributions of the neighborhoods are showing in a pie chart below, we can see that Maui has the highest percentage among all islands.

Chart, pie chart

Description automatically generated

Then I showed that the room type distribution among each neighborhood group below. We can see that room type of entire home/apt has the highest count. Again, Maui has the highest type of entire home/apt among all neighborhoods.

Chart, bar chart

Description automatically generated

I further analyzed the distribution of the price range per each neighborhood’s group shown below. I have set the price threshold to below $500 to narrow down my analysis. I would like to know the average and the price for each neighborhood group.

­

Chart, box and whisker chart

Description automatically generated

From the boxplot above, we can definitely observe a couple of things about the distribution of prices for Airbnb in Hawaii. We can see that Kauai and Maui has the high price range for the listings with about $220 as an average price. Honolulu has the cheapest average price among all.

I then analyzed the distribution of each property type shown below. As you can see, entire condominium has the highest percentage among all types, then follows by entire house.

Chart, pie chart

Description automatically generated

I also wanted to know about the review score ratings among all neighborhoods. Per graph below, we can see that Maui has the highest count of high review scores above 90, which is 4500 counts.

Chart, waterfall chart

Description automatically generated

Lastly, I have explored what the most used word in the listing name. The most used word could represent the selling value of their property for the prospective guests. First, I created a function to collect the words. I have used stop words library and also removed the words like numbers as they are referring the number of bedrooms. I then have selected the top 25 used words for listing names. Below is the distribution.

**Chart, bar chart, histogram

Description automatically generated**

Per results above, we can see that popular words are ocean, beach, condo, view, Waikiki, oceanfront, private and so on. Then I also generate a word cloud to visualize the popular words.

Text

Description automatically generated

# **Discussion and Challenges**

Overall, I am happy with my analysis. I am able to find out that Maui has the highest review scores and has the most Aribnb listings and the average price around $220. In conclusion, Maui is the most popular island among all islands in Hawaii. I will definitely visit Maui again. I have learned the data lifecycle and pipeline and also google cloud. In the end, I have shared my notebook into my github and anyone interested could apply this data pipeline into any city they want to analyze. Please see my notebook here <https://github.com/liushasky/hawaii-airbnb-reviews>

I did encounter few challenges at the beginning. I wanted to explore noSQL database on Google Cloud, so I used MongoDB Atlas on Google Cloud to analyze my Airbnb listing, shown below. After I set up the cluster, it turns out that I will have to pay data storage to store my Hawaii Airbnb dataset, so that I can run MongoDB queries for my dataset.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Then I decided that I could use Google Cloud compute engine to create VM instance to run Jupyter Notebooks. I was able to create an VM instance and then I started set up the environment for me to run Jupyter Notebook.

Graphical user interface, text, application, email

Description automatically generated

I installed pip3 and wget, and anaconda, and jupyter notebook.

Text

Description automatically generated

I then run Jupyter Notebook on the command line shown below.

Text

Description automatically generated

However, I did not have luck in getting the notebook open as I keep getting errors shown below. I have been debugging for hours and still getting the same error. I am guessing it is due to the environment set up, but with my limited knowledge on VM, I couldn’t figure it out.

Graphical user interface, text, application, email

Description automatically generated

Therefore, I decided that I can use AI Platform Notebook instance on Google Cloud to run my analysis, so that I won’t have to worry about the environment set up issue. Luckily, I was able to finish the project and the process has been smooth on AI Platform Notebook instance. I will continue working on the VM to see how I can get it work.

# **Conclusion**

In this project, I have analyzed the Airbnb listings dataset for Hawaii, including the neighborhood groups, property type, room type, price, and listing review. I found out that Maui has the highest listing reviews and review score among all neighborhood groups. I was able to use AI Platform Notebook instance on Google Cloud to run my analysis. I was able to perform each step are list in the USGS lifecycle model [1], which includes Plan, Acquire, Process, Preserve and Publish/Share Airbnb listing analysis.

# **Reference**

[1] Plale, Beth, and Inna Kouper. “The Centrality of Data: Data Lifecycle and Data Pipelines.” *Data Analytics for Intelligent Transportation Systems*, Elsevier, 7 Apr. 2017, www.sciencedirect.com/science/article/pii/B9780128097151000043?via%3Dihub.