Final Project Report CS327E

Robert Saenz rs42744

Stefanos Kapetanakis sk35337

**Introduction**

**Main Goal**

The purpose of the course, CS327E Elements of Databases, was to create a data warehouse and pipeline through the conglomeration of several datasets/databases together into a larger whole, while also gleaning useful data from this warehouse using SQL queries, and data visualizations. In particular, we conducted a data-driven analysis of Airbnb data in Austin and addressed the revenue being generated, the occupancy rate, the amount of rentals available, and how these metrics have changed over time.

**Datasets/Databases**

The datasets that we were instructed to use were based on housing, with long term and short term data being available. First off was the Airbnb rental data for three different cities, those being Austin. Portland, and Boston. Each of these three cities had 10 tables associated with them, containing information like listing prices, neighborhoods of listings, and host information for each host that uploaded a listing. Eventually joined with this data was the *Zillow* database information, containing the monthly prices for houses listed on *Zillow*’s website, where each kind of apartment/house was classified as 1 to 5+ bedroom, and organized into the appropriately named file based on this information.

Overall there were 36 files that were being interpreted and accessed in the BigQuery data warehouse..

**Software and APIs**

The tools used during this course were largely based on Google’s GCS (Google Cloud Services) tools. These include Google Cloud Storage for holding .csv files for upload and storing data about Dataflow jobs (which were used for data pipelining in place of MapReduce). Another major tool was BigQuery, the actual big data warehouse tool that was used to store and access the previously mentioned data files. These files were accessed using SQL, as a way to easily query them, and were finally visualized for easier consumption using Google’s Data Studio visualization tool. Apache’s Beam API was used alongside BigQuery and Dataflow to easily pipe data into BigQuery. Lastly Git was used for easy control of project files and bug/problem tracking for each of our labs, and UT’s Stache was used for easy password access among partners and teaching assistants/professors.

**Analysis of Data and Processes**

**Initial Setup**

The initial setup involved creating a GitHub for version control between multiple people. We did not extensively use GitHub to work remotely, but rather for easier collaboration in person so that we could simultaneously improve the project at the same time. Most of our final milestone files were placed in relevant folders within the GitHub, where we turned in and submitted them for each milestone in a weekly fashion. From here we also worked cooperatively through Google’s GCS implementation where we could both view and query the necessary files for the project using the IAM (Identity Access and Management) feature.

**Database Design and Schema Information**

We were handed 5 files in the Zillow dataset to use, and to incorporate among our 3 Airbnb city datasets. Our Airbnb datasets were organized by city for easier classification and interpretation of data so that we could easily compare among the three geographic locations. Each of these was populated from earlier labs in a normalized form so that data manipulation and access was as easy as possible, and so SQL query generation was as simplified as we could possibly make it for easier interpretation of the statistics for each city, especially for data regarding Austin long term and short term rental data.

In order to link the Airbnb and Zillow datasets for more information about Austin, we created a new SQL view that was intended to link together the 2 tables for optimal visualization and comparison. This view was called the “Revenue Crossover Point” and was extremely useful for gleaning information about the tables. This data includes connections on the number of Airbnb rentals available, the information about the change of this metric over time, and the impact of Airbnb earnings on long term rentals in Austin (represented by proxy through the Zillow data).

In this view, we joined the Airbnb and Zillow datasets on the bedroom, zipcode, and date fields. Then we computed the revenue crossover point metric, which was equal to the ceiling of Zillow’s median rental price per month over Airbnb’s median rental price per day. This metric we computed represents the amount of days a month an Airbnb host would have to rent out their property in order to receive the same revenue from a median long-term rental.

**Data Analysis (SQL Queries)**

Once we joined all of the data together we wanted to create some metrics in order to figure out and visualize how many short-term rentals are available, what their occupancy rate is, how much revenue is being generated over time, and how these metrics changed as time went on specifically for the city of Austin. In order to create these metrics we made 3 different SQL queries that were relevant to each metric. We explain in detail how we created these metrics below, and the SQL queries for these 3 metrics are all shown at the end of this report in the appendix under the title “SQL Queries / Metrics”.

In order to figure out the revenue generated for the rentals, we multiplied together the listing price, minimum nights, and number of reviews for each rental and labeled it as a new variable called revenue. We also created a column for the city and selected the date so we could then order the results by the date for each listing. It’s also important to note that we joined together the Listing and Calendar tables so we could use the date from the Calendar table with these other variables. This ORDER BY statement would allow us to display the changes in revenue over time.

For our second metric, we wanted to find the number of short-term rentals that were available in Austin and show the change over time. In order to do this, we use the COUNT() function on the listing id and labeled this as num\_of\_rentals. This would count the amount of different listings there were. In the select statement we also added the city as Austin and the date. We also joined together the Listing and Calendar tables so we could use date to order the results. Ordering the results by date allowed us to show the number of rentals for each date as time went on.

For our last metric, we wanted to find the occupancy rate of the rentals. In order to estimate the number of days that were occupied for each rental, we multiplied together the minimum nights and number of reviews for each listing. This would give us a good estimate of the number of days each rental was occupied. This is where we ran into a problem. We couldn’t figure out how to show this metric over time, since the Calendar table has multiple dates for each listing. So when we tried to group by or order by date, the query would not work. We tried many different solutions, but in the end could not figure out a proper way to write and display this metric in SQL and BigQuery. The SQL for this query is in the appendix below.

**Visualizations**

With the metrics we created as explained earlier, we wanted to make visualizations of them so we could get a better understanding of the output. So we input the queries into BigQuery in order to visualize them, and the visualizations are posted below in the appendix.

Our first metric was revenue generated over time for the listings, and the visualization for it is below titled “Revenue Generated Over Time”. ­Looking at the graph you can see that the average revenue has actually lowered from 2015-2018. There was a slightly big drop after 2015-10, but the revenue rose up slightly in 2016-06, but still not higher than before 2015-10. From there, the revenue stayed at a relatively steady rate.

Our second metric addressed the number of rentals available over time, and our visualization of it is below titled “Number of Rentals Available Over Time”. It’s important to note the reason for the drop to 0 between Oct 16, 2016 and Feb 21, 2017 is because the data had no listings during that time due to some unknown error. Looking at the graph though, you can see that there is a growing trend in the number of rentals available as time goes on, and the amount of rentals nearly doubles from May 25, 2015 to Feb 21, 2017. After Feb 21, 2017, the number of listings has stayed relatively consistent.

Our third metric, the “Avg Daily Airbnb price vs Zillow monthly price” visualization displayed a side by side view of the average cost for Austin in each of the two categories, with the dimensional axis having time divided into monthly intervals from May 2015 to Jan 2018. As can be seen from the graphs themselves, there does appear to be a minor effect of correlation between the two variables. The general trend is a slight rise in prices for both over time from 2015 to 2018, though it is by a relatively small margin. I take this to mean that while there may be a bit of a linear relationship between the variables, it is too slight to be confident, and so may need further in depth study over a longer time period in order to say for sure that the effect conclusively exists and is not the result of statistical noise.

**Conclusion**

Throughout this course and project we learned how to effectively create a data pipeline from several data sets, visualize the database with the ingested data using ERDs, implement and create metrics using SQL in BigQuery in order to visualize them, and analyze the data to infer valuable insights. With this came many challenges along the way, but this only allowed us to improve our debugging and analytical skills. Our main focus for this project was to dig deeper into the given Airbnb data for the city of Austin with a data-driven analysis by exploring the revenue generated over time, the occupancy rate over time, and the amount of rentals available over time.

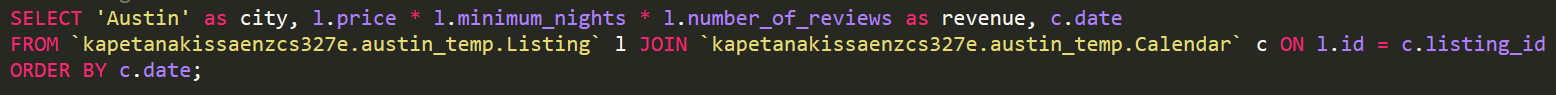
We initially setup our database using Google’s GCS, Postgres, and Github, which were all invaluable tools for this project. Github was extremely useful in allowing us to share our files, code, and other relevant work, allowing use to easily work in collaboration. We ingested the Airbnb data into the GCS, and later did the same with the Zillow dataset in order to draw comparisons between the two. We created a view called the “Revenue Crossover Point” which really let us gain some insight into the comparison between Zillow and Airbnb’s prices.

Once we had the database set up, we dived into some analyses by creating relevant metrics with SQL and visualizing them in BigQuery. The first metric we created and visualized was the revenue generated from the Airbnb rentals. From the visualization we inferred that while the average revenue has taken a slight dip over time, it still remains at a relatively high and constant rate. For the second metric, we addressed the number of Airbnb rentals available over time in the city of Austin. Since the city of Austin has been growing in popularity over the last few years, we expected that the number of rentals would raise over time, and that’s exactly what our results were. The surprising part of this though was that there was a span from about June 2016 to October 2016 where the number of rentals actually decreased. This could be due to the error we mentioned earlier about the data, where the was no data from October 16, 2016 to February 21, 2017. For our third metric, we were unsuccessful in properly visualizing the data, but we did figure out how to calculate the estimated days occupied for each listing.

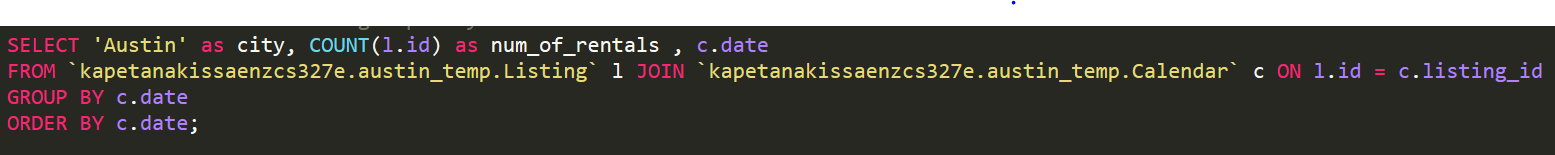
**Appendix**

**SQL Queries / Metrics**

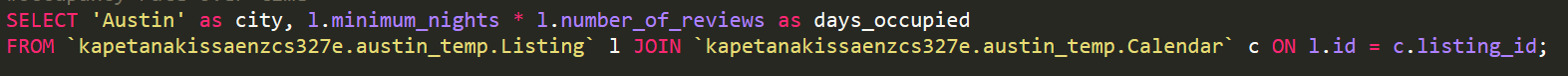
Revenue Generated Over Time:



Number of Rentals Available Grouped by Date:

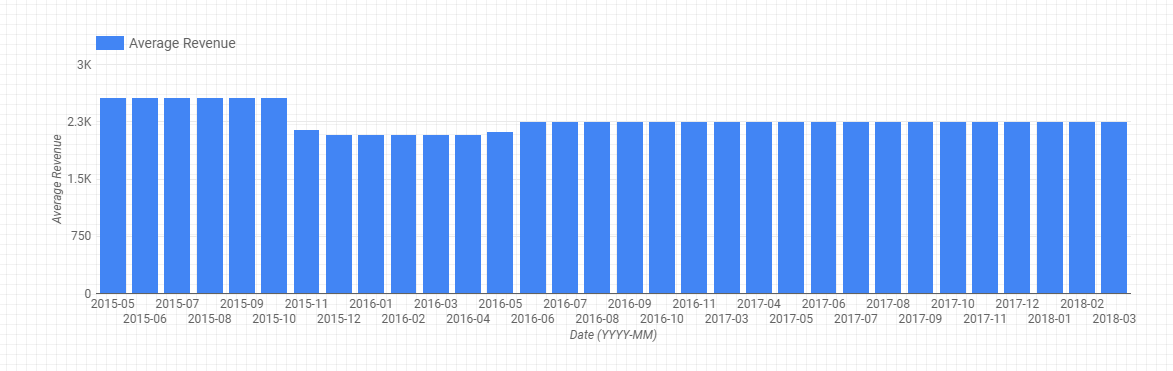


Occupancy Rate:

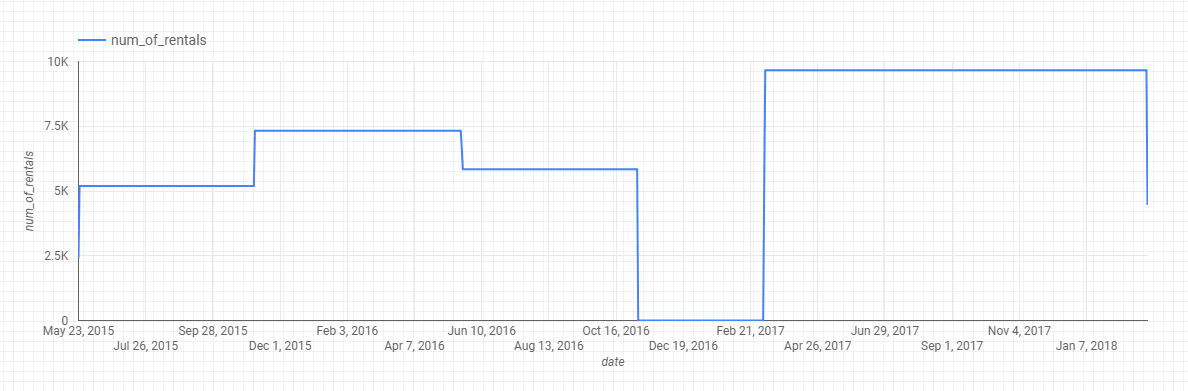


**Visualizations**

Revenue Generated Over Time:



Number of Rentals Available Over Time:



Avg Daily Airbnb Price vs Zillow Monthly Price:

