**Data exploration on lodging market in Oswego County**

Having learned statistics together in many classes, we decided to work on the capstone project together. Talking with our advisor, Professor Zheng Hao about this decision, he provided us a topic to work on. Spending more than a semester to work on this, we came up with an analytics on data exploration on the lodging market in Oswego County.

Firstly, the data package given to us was provided by Oswego County Community Development, Tourism, and Planning. It included 9 different files: 4 excel files, 4 raw data files, one Oswego Zip Codes file. The 4 excel files were work from a third party company so we shifted our focus to the 4 raw data files. Each file had a time frame of two years. The time period was overlapped in each file but fortunately there were no missing observations, but rather a lot of duplication. Therefore, data organization played a vital role beforehand. We combined 4 data raw files into one single file by removing duplicates.

After combining all the files into one file, we got a data set containing 371 observations and 34 different variables. The statistics were recorded from 2018 to 2021. On each day, for each variable, there were 5 regions with Oswego County being the whole region and the other 4 Lake Shore, Oneida Lake, Salmon River, Tug Hill being the subregions. For each subregion and region, there were statistics from two providers: AirBnb and HomeAway. One thing to note is that the sum of 4 subregions may be larger than the whole Oswego region due to some overlapping. For example, a room was counted more than once.

When receiving the topic, we were asked to measure the amount of visitation these areas received through the owner rental platform. The hints were monthly traffic for each region including strongest and weakest months, year-to-year trends, length of stays, and revenue collected as rental income. Based on what was provided, we came up with the following adjustments. We decided to analyze data from 2019 as we were unable to gather sufficient data in 2018, making it impossible to view the 2018 monthly trend. For length of stay, we added two more columns: Nights per Stays and Room Nights per Stay. Finally, we chose to analyze only the Airbnb provider since we were unable to verify the HomeAway one. The following step was to put everything in R to make needed graphs and plots.

The two of us had little experience in R. We only did some homeworks using the software in other classes. Professor Zheng provided us with sources to read and learn by ourselves. The most useful thing we found was that he gave us exercises and required us to present what we got in every single meeting, which forced us to learn and master the software. To make things easier, we broke the data set file into some sub data sets. Then we imported each file into R to make bar graphs and time series plots. We plotted nearly 100 graphs and chose some to report on based on the required goals.

Having finished graphing and plotting, we moved on to commenting on trends and seasonalities. To answer the first question, monthly market supply and demand, we utilized three different variables. The first one was available nights, meaning the amounts of nights that were available to be booked. Night booked was the second variable, counting the total of guests occupied nights. Last but not least, adjusted occupancy rate stood for the percentage of available nights that had been booked. This rate could be calculated by taking Available Nights divided by Night Booked. The first graph illustrated the monthly average Nights Booked (in blue) versus Available Nights (in neon color) in Oswego County. This graph showed that Oswego County had averagely at least 4000 available nights each month. The highest month was November with 5693 nights, while the lowest was February with 3811 nights. The number of nights booked was picked in August with approximately 2500 nights, accounting for 51% adjusted occupancy rate. The least busy month was January with only 16% of booked nights. Putting this graph side-by-side with the monthly Nights Booked versus Available Nights, we could see a similar trend. November still had the highest availability and August was the busiest month in each year. Two subregions Salmon River and Lake Shore had quite similar trends as Oswego County had. The highest number of monthly average available nights were 2101 nights and 3200 nights, respectively, recorded in November. February still had the least open nights of 1342 nights in Salmon River and 2142 nights in Lake Shore. The busiest months in both subregions were August with the occupancy rate of 50% and 48%. On the other hand, January had least booked nights with just less than 20% of occupancy rate. Same as Oswego County, the graph of monthly Nights Booked versus Available Nights in these regions also had the same patterns as the average monthly plots had. The last two regions, Tug Hill and Oneida Lake had the highest availability of 3000 nights and 855 nights in November and April, respectively. The most booked nights were recorded in August of 1279 and 410 in each region. February in Tug Hill had least availability but the least booking was January of less than 20% of occupancy rate. While in Oneida Lake, July had the lowest available nights of just 606 nights, but January had the lowest occupancy rate of 24%. When looking carefully, we saw that Oneida Lake had a strange pattern compared to other subregions. The graph of monthly Nights Booked versus Available Nights also showed that. For some months, the number of availability and booked nights in 2019 are higher than those of 2020 and 2021, specifically in the second half of each year.

With the first question being answered, we turned to the second one, which was the year-to-year market trend. For Available nights, which was the amount of nights that were available to be booked, Oswego County had available nights with approximately 68720 nights available. The number started at 44887 available nights in 2019, then increased to 58109 nights for 2020, then finally increasing to 68720 available nights for the year 2021. Oneida Lake region had the least amount of nights available with only 9374 nights available for 2019, 9373 nights available for 2020 and then a huge decrease to 8886 nights available for 2021. Tug Hill was seen to have a similar pattern as Lake Shore. Salmon River had 15889 available nights for 2019, 19997 available nights for 2020 and finally jumped up to 28320 available nights for 2021. In general, despite the Covid, the number of available nights has been increasing from year to year except for the Oneida lake region. In terms of Nights Booked, it is no surprise that Oswego County booked nights increased from year to year. It started at 12415 nights booked in 2019 then went up to 18339 nights booked for 2020 then finally increased again to 22698 nights booked for 2021. Lakeshore, Oneida Lake, Salmon River, Tug hill all followed the same pattern as the graph for available nights. Oneida Lake had the least nights booked with only 3650 nights booked for 2019, 3588 nights booked for 2020 and finally 8886 nights booked for 2021. In general, yearly nights booked followed the same pattern as available nights for each of the regions. About occupancy rate, four regions: Oswego County, Lake Shore, Salmon River, and Tug Hill had quite similar numbers in the 2019-2021 period. The number fluctuated between 24.74% and 33.96%. Oneida Lake had a high occupancy rate with 39.56% in 2019, 39.66% in 2020, and 2021 was 39.00%. These groups of numbers suggested that regions Lake Shore, Salmon River and Tug Hill had similar behaviors; these three regions mostly determined the big picture for Oswego County.

The third topic to be analyzed was length of stay. We chose two variable plots: average nights per book and average room nights per book. Average nights per book was the average number of unique nights per booking within the reporting period. On the other hand, average room nights per book was the average number of room-nights booked within the reporting period. Looking at the time series graph, Oswego County's average nights per book had its highest average night per book in the month of March with about 7.835 nights and its lowest month was in October with about 3.175 nights per book. It was important to note that this graph was the monthly average average nights per book for Oswego County. The next graph showed the monthly average nights per book from 2019 to 2021 in Oswego County. March of 2019, March and April of 2020, March and December of 2021 outnumbered every other month. Moving on to the time series plot of other region’s monthly average nights per book, we saw that Lake Shore was higher than the other regions with March being the month that had the highest average nights per book with 8.90 nights per book. Salmon River had the least average nights per books of all the regions. October had the least average nights per book recorded for Salmon River with approximately 2.97 nights per book. The graph of other region’s monthly average nights per book also showed that March and December were the busiest months in each year of each region. For room nights per book, Oswego County’s highest months were March and April with 13.79 nights for March and 14.831 nights for April. The lowest average room nights per book occurred in the month of October with 6.758 room nights per book. The difference between average room nights and nights per book was simply that the number of rooms booked was higher than the number of nights booked. The Oswego County’s monthly average room nights per book also illustrated that April and December in each year had a higher number of booked room compared to other months. When looking at the other region’s average room nights per book, Oneida Lake had the highest average room nights per book compared to other regions. Oneida Lake had April as ist highest month with approximately 17.11 average room nights per book. Lake Shore had the lowest average room nights per book of all the other regions. Lake Shore had its lowest month in October with its average room night per book at approximately 5.88 nights. The time series graph of the other regions’ average rooms night per booked showed that April of 2020 was the highest month in each region during the 2019-2021 period.

Finally, the monthly and yearly revenue was illustrated. The variables being used included gross revenue being the sum of transactions for all stayed nights, adjusted revenue per available listing, and adjusted revenue per room. Oswego County’s highest monthly gross revenue was noted in August with roughly 394428 thousand dollars. While the lowest month was January having only 67542 thousand dollars, we could see a huge range between the most and the least earned. Among the four subregions, Tug Hill had the highest revenue. This number was recorded in October with 245 thousand dollars. Salmon River and Oneida Lake also had the peak revenue in October with 232 thousand and 198 thousands dollars respectively. Lake Shore also had the similar peak revenue as Oneida Lake had, except for this number was earned in August. The weakest earning months for all of these subregions were in January with Tug Hill having 61 thousand, 56 thousand in Lake Shore, Salmon River was 32 thousand, and the lowest was Oneida Lake with 22 thousand dollars. For monthly adjusted revenue per available listing, the weakest number fell in January for all of the regions. This number was ranging from 442 dollars to nearly 1000 dollars. Some subregions even outnumbered Oswego County such as Oneida Lake had more than 3000 dollars and Tug Hill had 2520 dollars compared to 2510 dollars of Oswego County. Lake Shore and Salmon River had lower revenue per listing than Oswego County but the number was not far away with 2173 dollars and 2499 dollars, respectively. Next, we analyzed monthly adjusted revenue per available room. Looking at the graphs, each region and subregion had quite similar highest numbers with approximately 40 dollars per room recorded in August and October. The least earning month was still January with the number ranging from 7.5 to 15 dollars per room and Lake Shore had the lowest revenue among five regions and subregions. We noticed that this number was much smaller than the monthly adjusted revenue per listing that we looked at above. The reason for this was that this was the average earning for each room, and one listing might have more than one room available to be booked. Turning to yearly trends, we saw the total gross revenue increase year-by-year. For Oswego County, in 2019, the revenue recorded was approximately 1.6 million dollars. Then it went up sharply till nearly 2.7 million dollars in 2020 before reaching the peak of roughly 3.7 million dollars at the end of the 3-year period. Other subregions followed a similar upward trend. For average revenue per listing, the number also increased by year despite the presence of Covid. Interestingly, Oneida Lake outnumbered other subregions, and even the big region Oswego County. In 2019, Oneida Lake earned 17 thousand dollars, compared to less than 15 thousand dollars of other regions. Similarly, 2020 and 2021 of Oneida Lake also had a few thousand higher incomes than the rest. For yearly average earnings per room, the trend was identical to the yearly average revenue per listing. The number was quite close in each region, approximately 200-260 dollar per room in 2019, 250-350 dollar in 2020 and 300-360 dollar per room in 2021.

To conclude, from the graphs, we saw that even though COVID was present, the number of bookings and the revenue still increased compared to the previous years. That was, it went up from 2019 to 2020 to 2021. The month that had the most bookings was August, while the lowest month was January. For some reasons, March also had higher bookings compared to other low-booked months. April was the month with the highest nights and room nights per book while October was the month with the least nights in this scenario. The highest revenue was seen from the months of June to October.

After spending nearly two semesters doing this project, we learned a lot more than what we expected. Before doing this capstone project, we barely knew anything about R. Under the thorough instruction from Professor Zheng, we could import the .csv file into R, define variables, draw bar plots and time series graphs. Not only that, we understood the process of organizing and cleansing data. The most important part after all was to know how to comment on the graphs, find out the trend, the seasonality, and also to combine and report everything in a complete presentation. We would like to thank the SUNY Oswego and its Faculty for providing us with the materials and knowledge to complete the Capstone.

**Appendixes**

Figure 1Chart, bar chart

Description automatically generated

Figure 2

Figure 3

Chart, histogram

Description automatically generated

Figure 4

Chart, bar chart

Description automatically generated

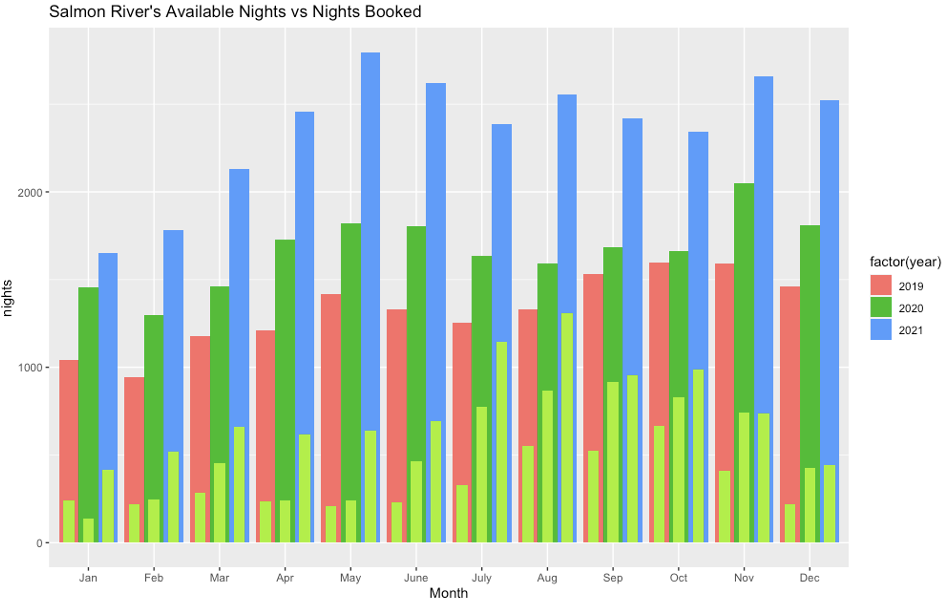
Figure 5

Figure 6

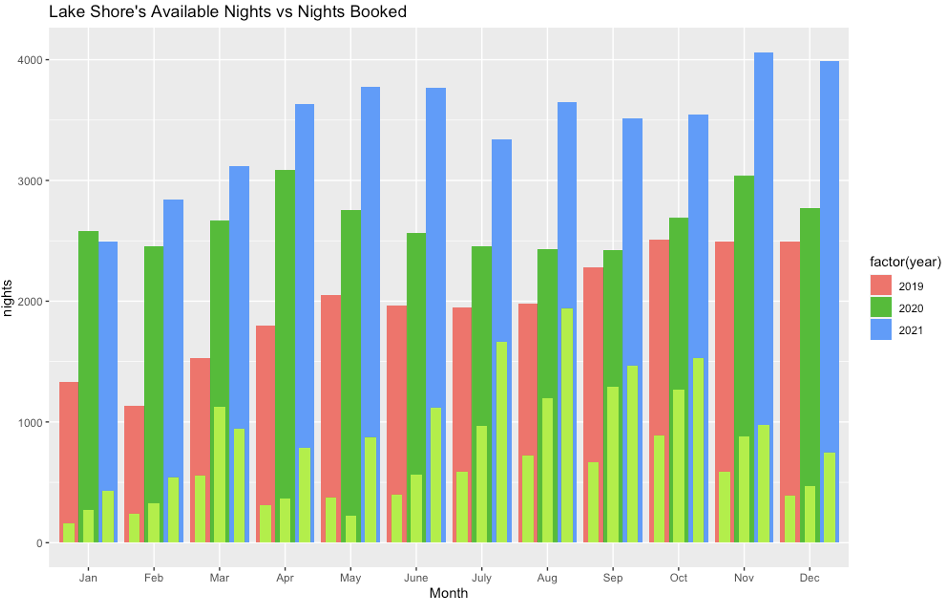


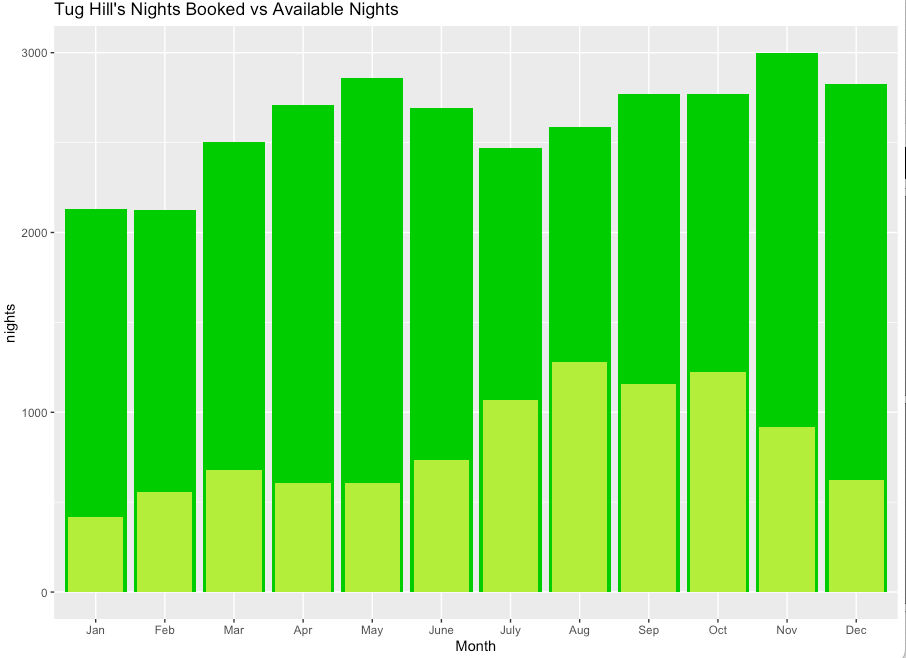
Figure 7

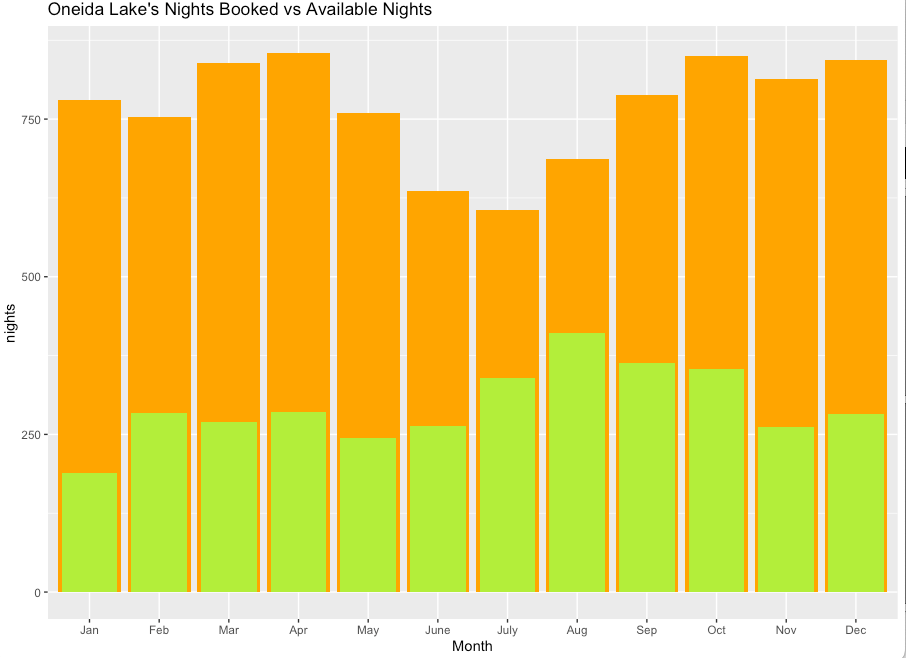
Figure 8

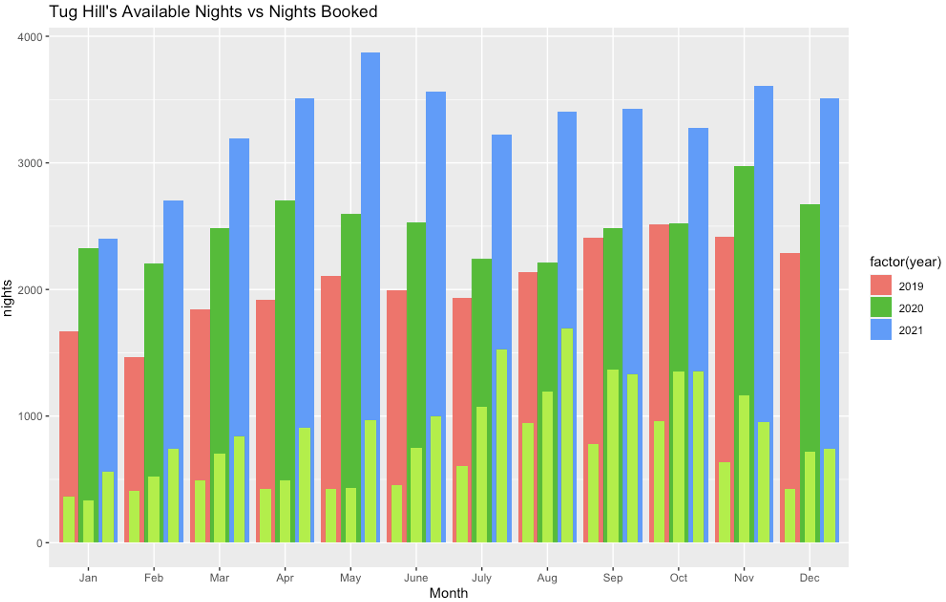
Figure 9

Figure 10

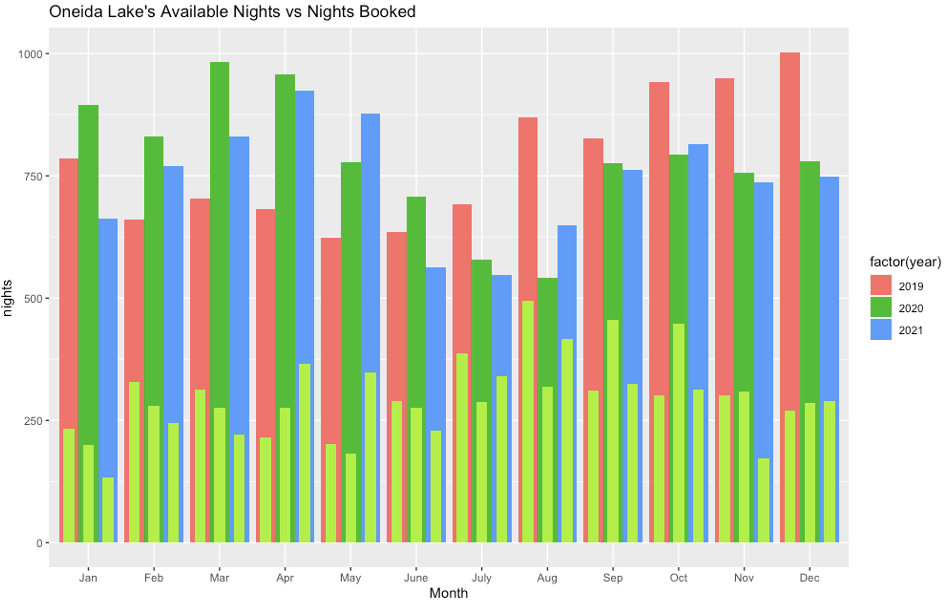


Figure 11Chart, bar chart

Description automatically generated

Figure 12

Chart, bar chart

Description automatically generated

Figure 13

Chart, bar chart

Description automatically generated

Figure 14

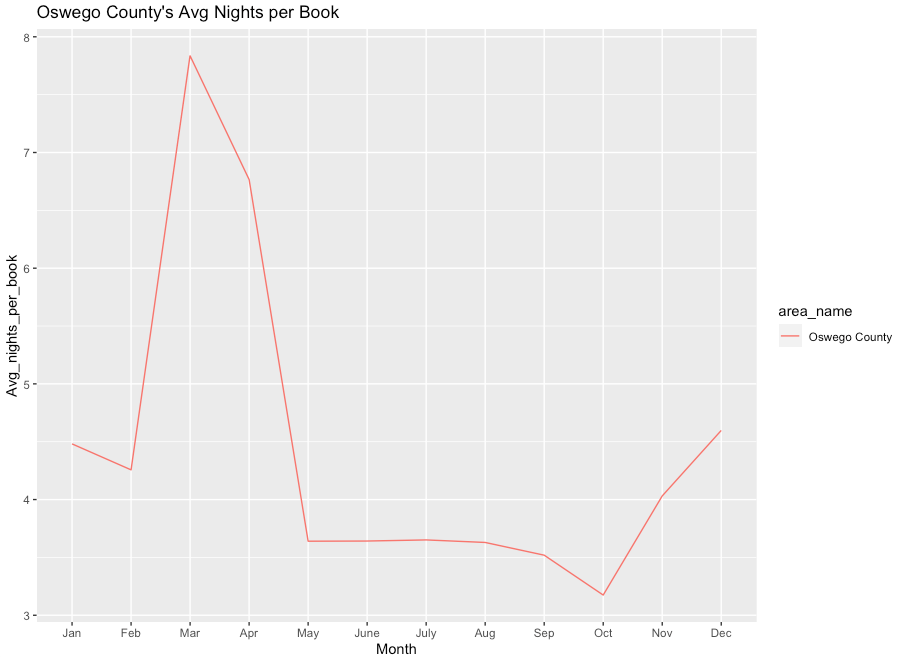


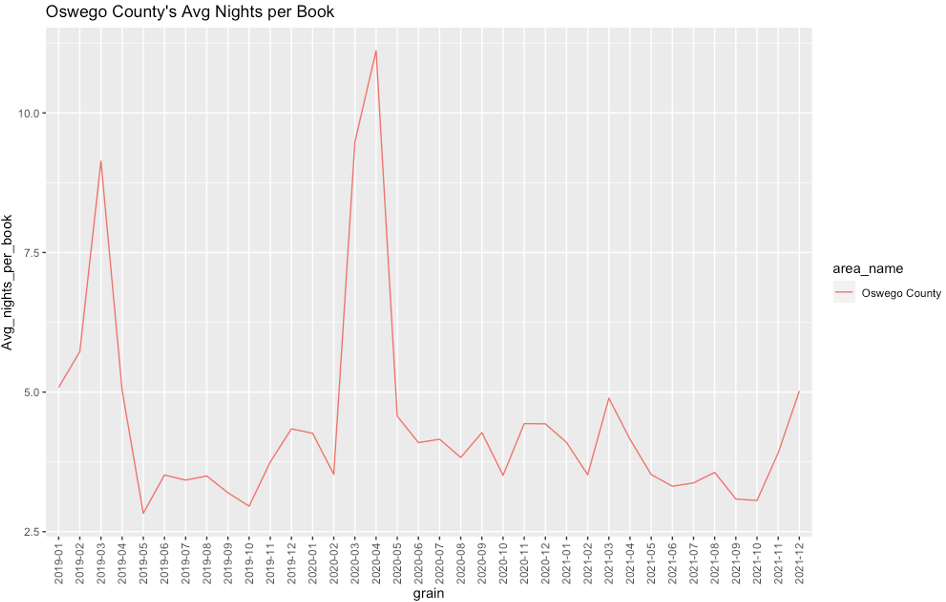
Figure 15

Figure 16

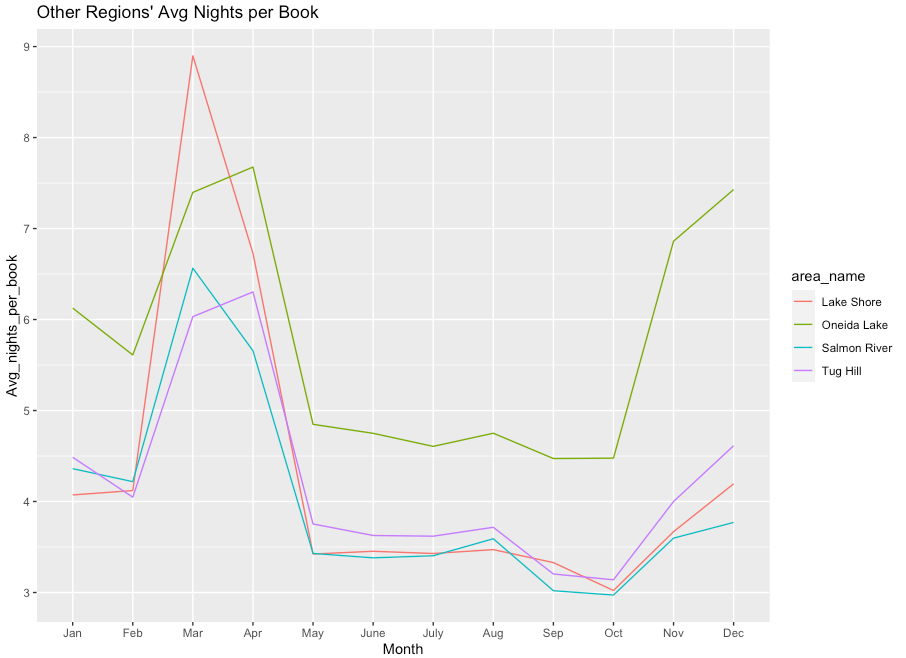


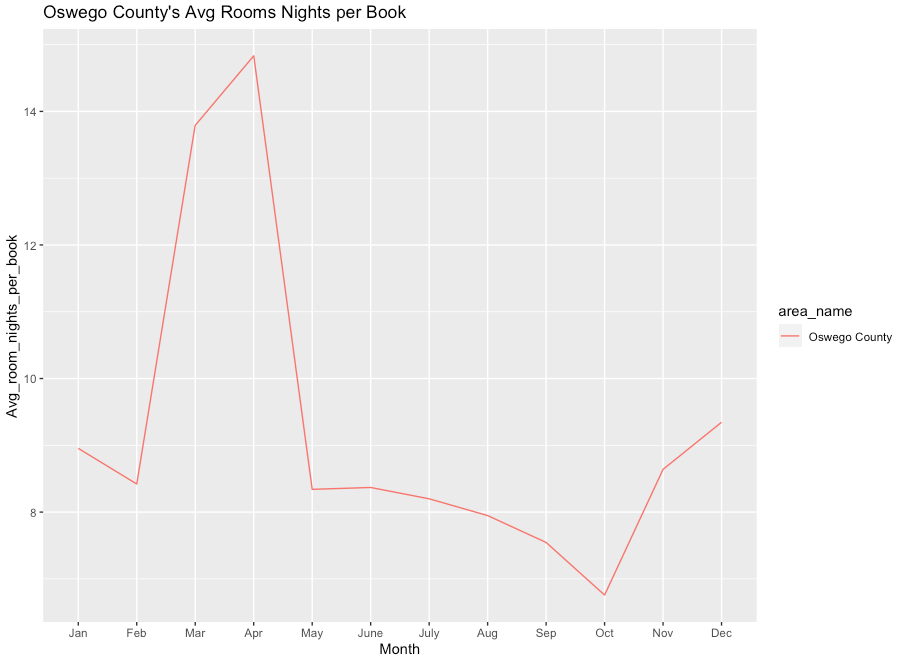
Figure 17

Figure 18

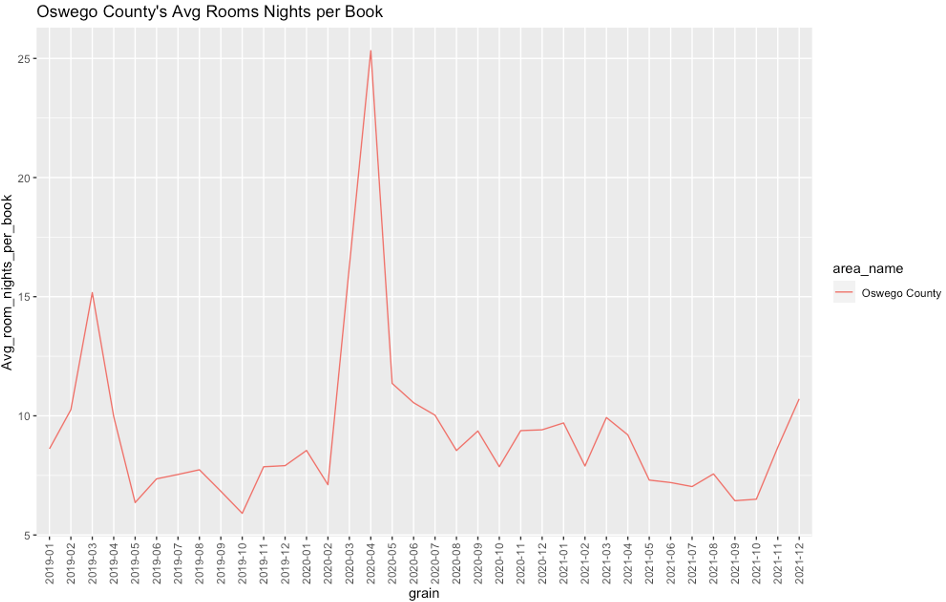


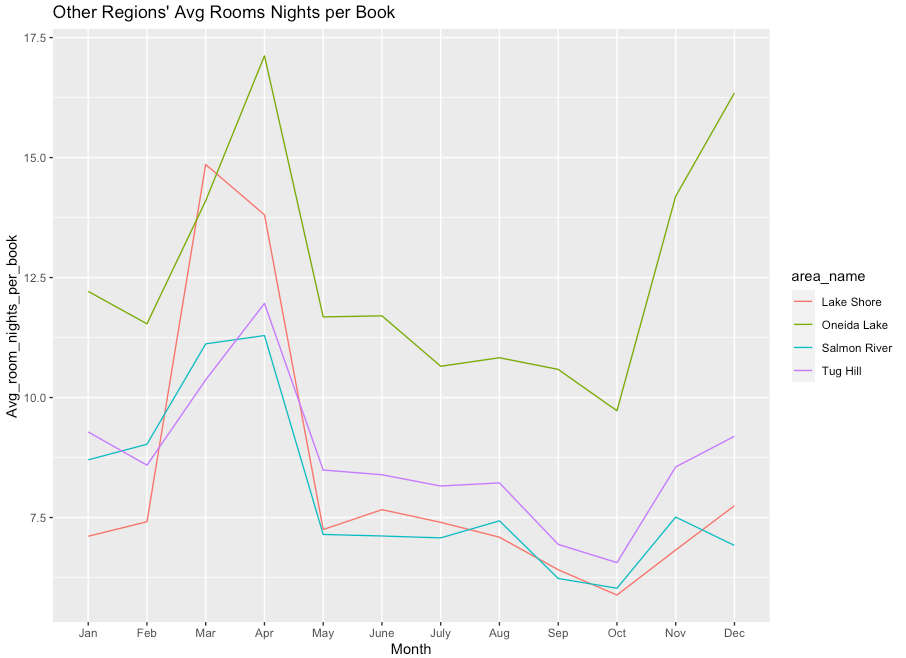
Figure 19

Figure 20

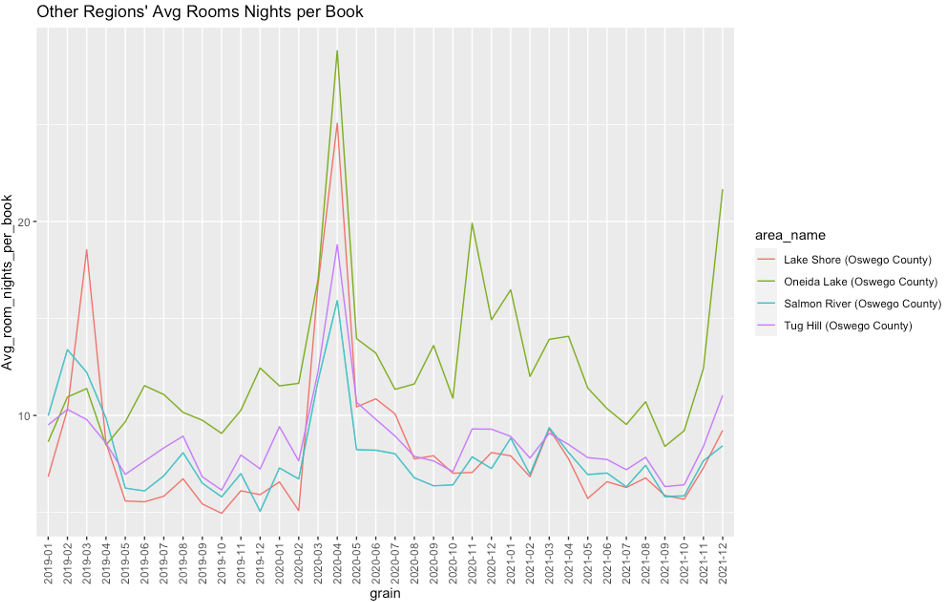


Figure 21

Figure 22Chart, histogram

Description automatically generated

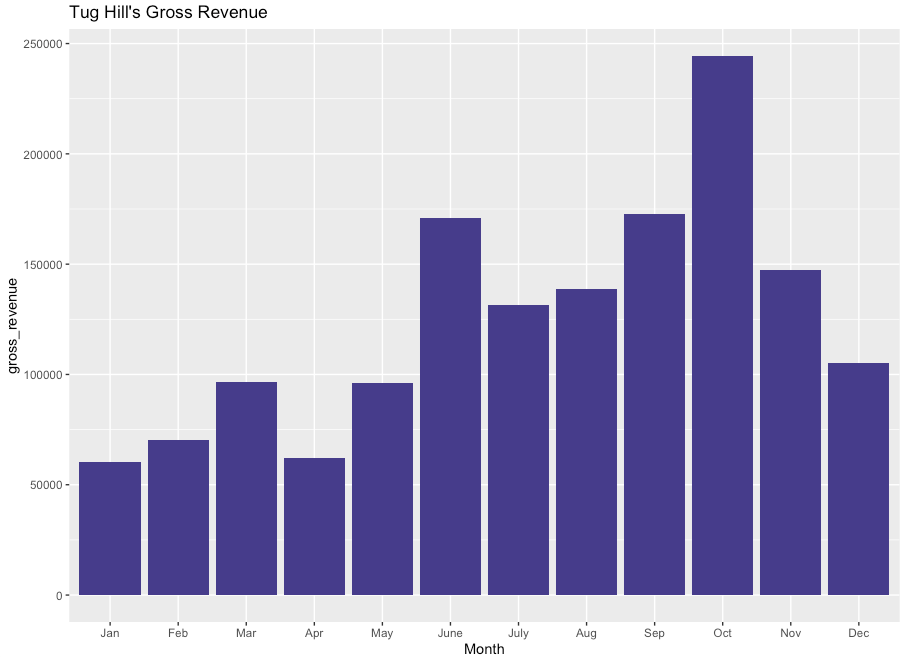


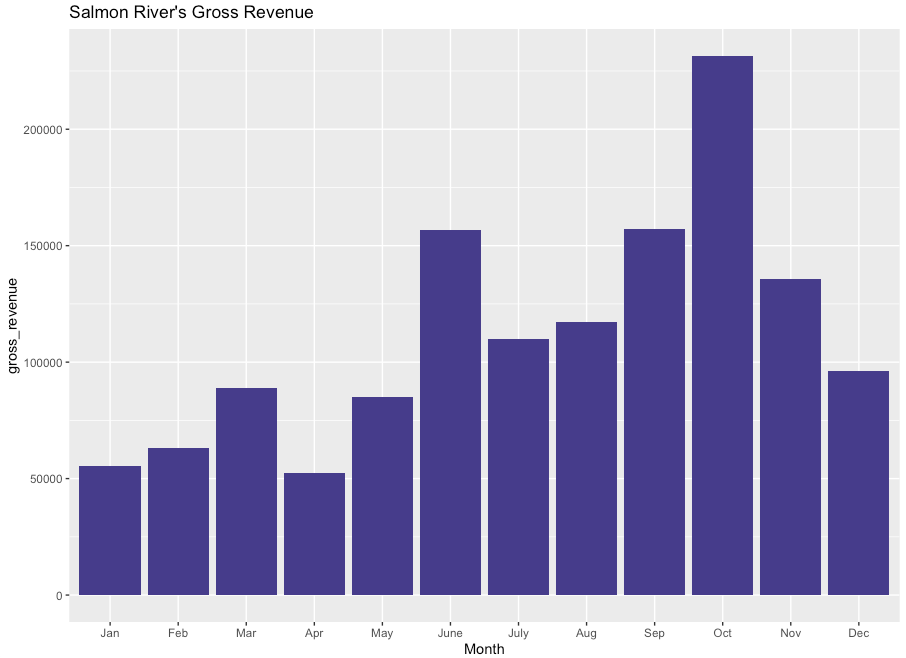
Figure 23

Figure 24

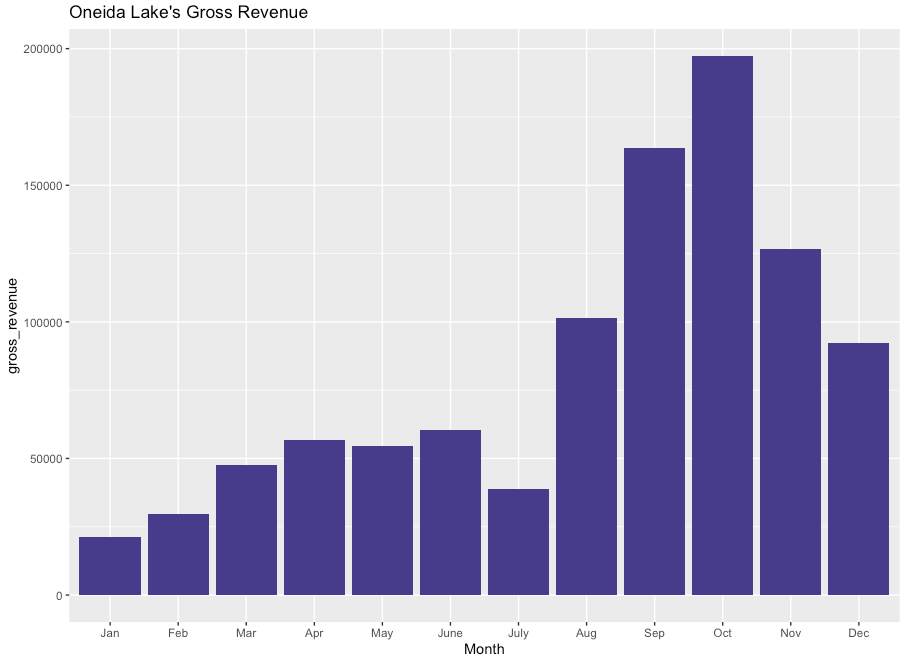


Figure 25

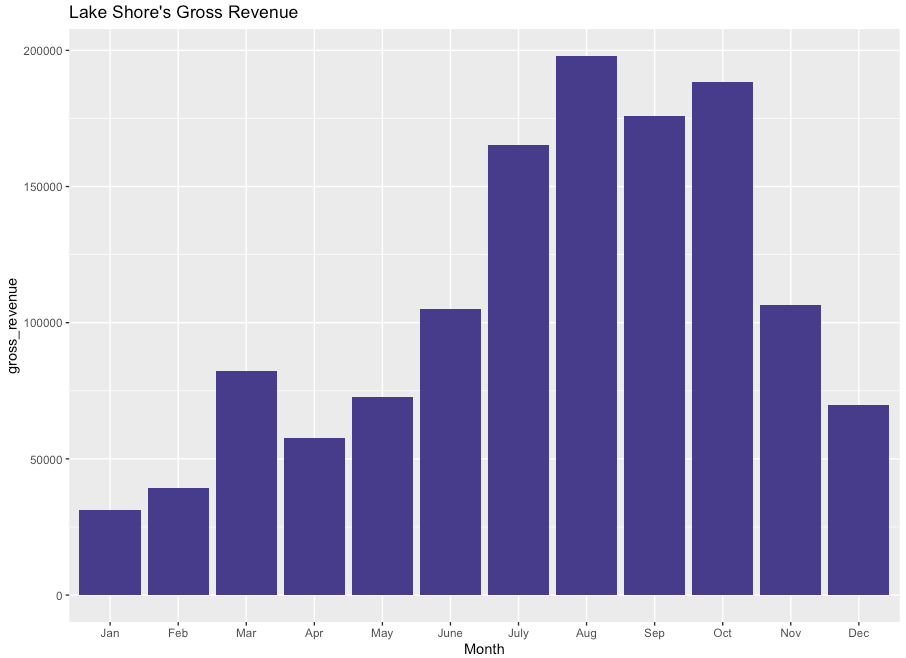
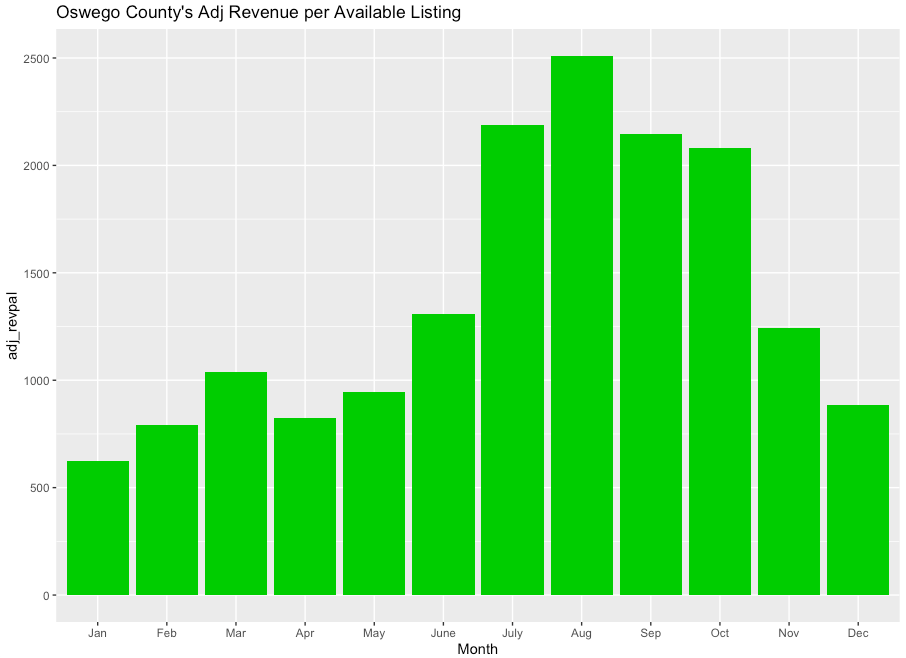
Figure 26

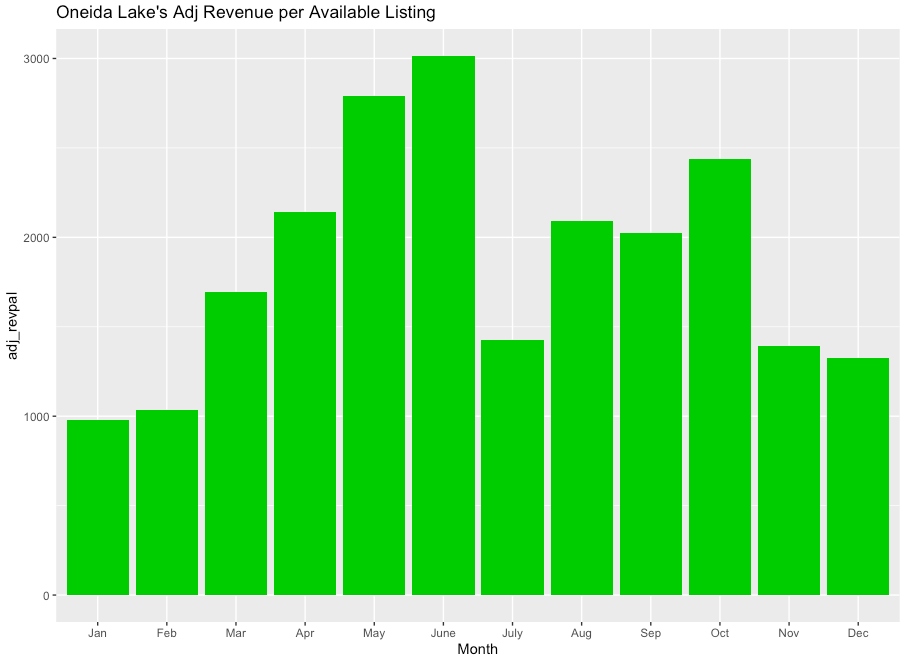
Figure 27

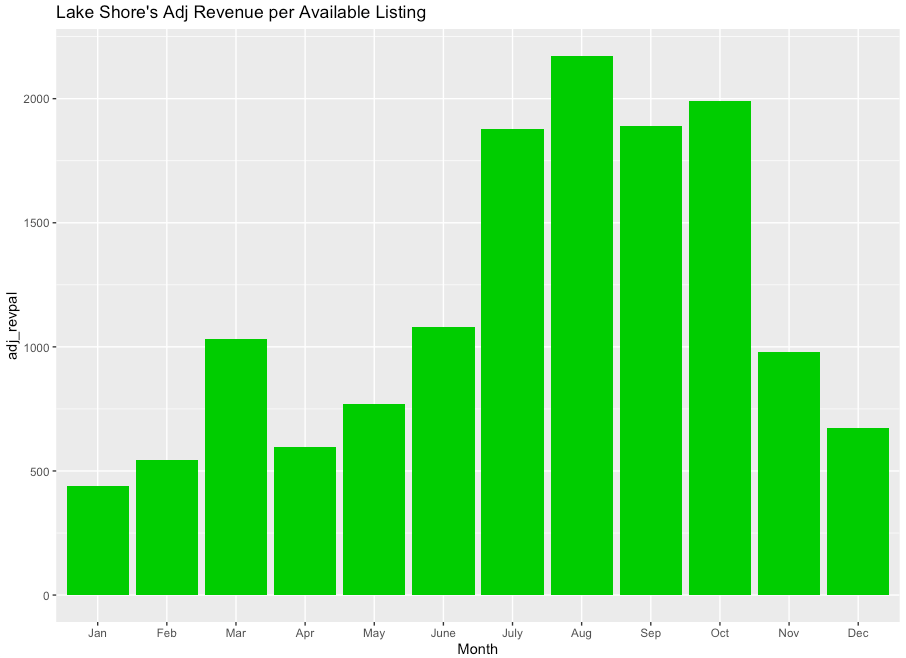
Figure 28

Figure 29

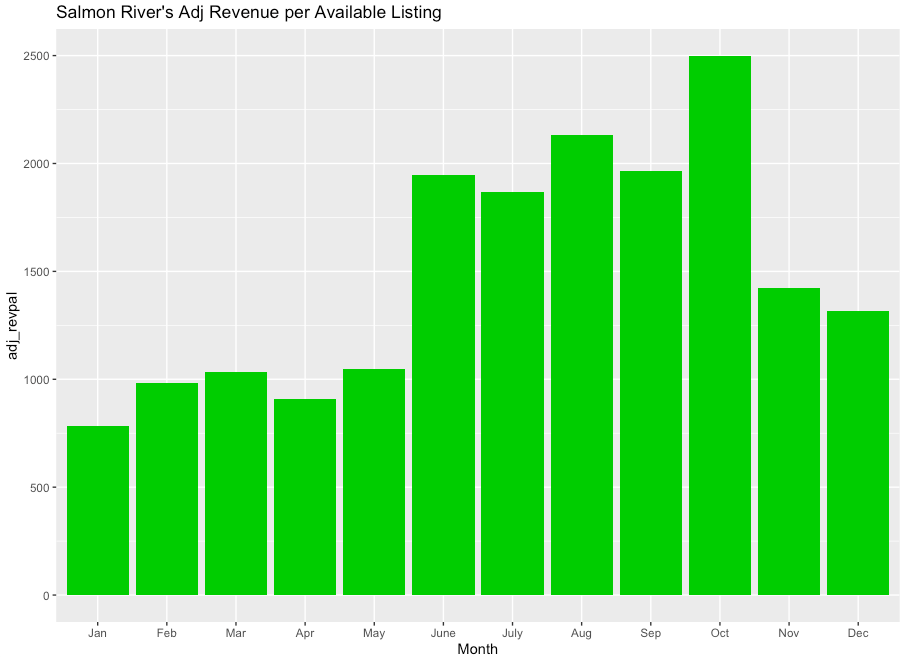
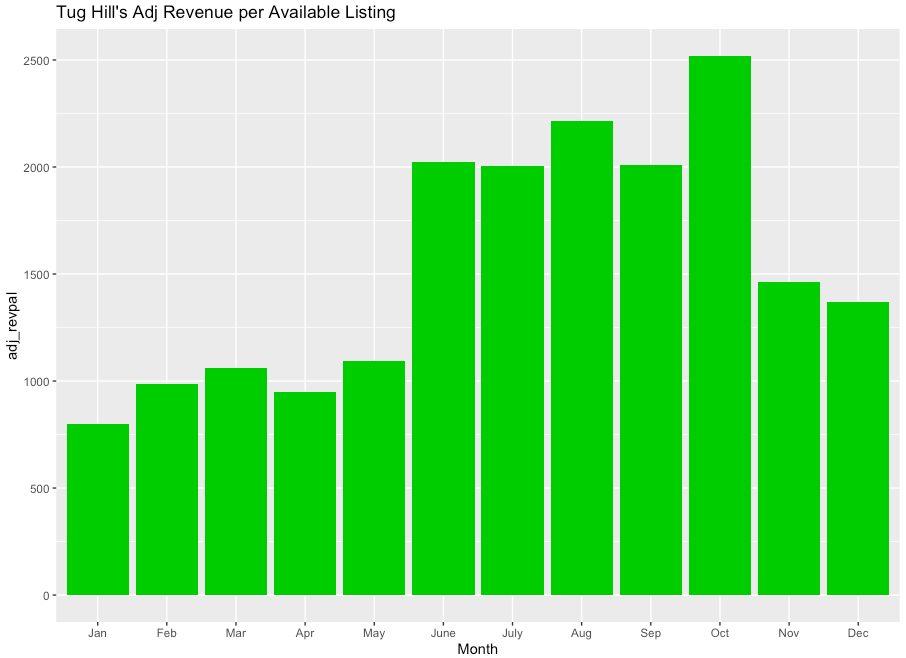
Figure 30

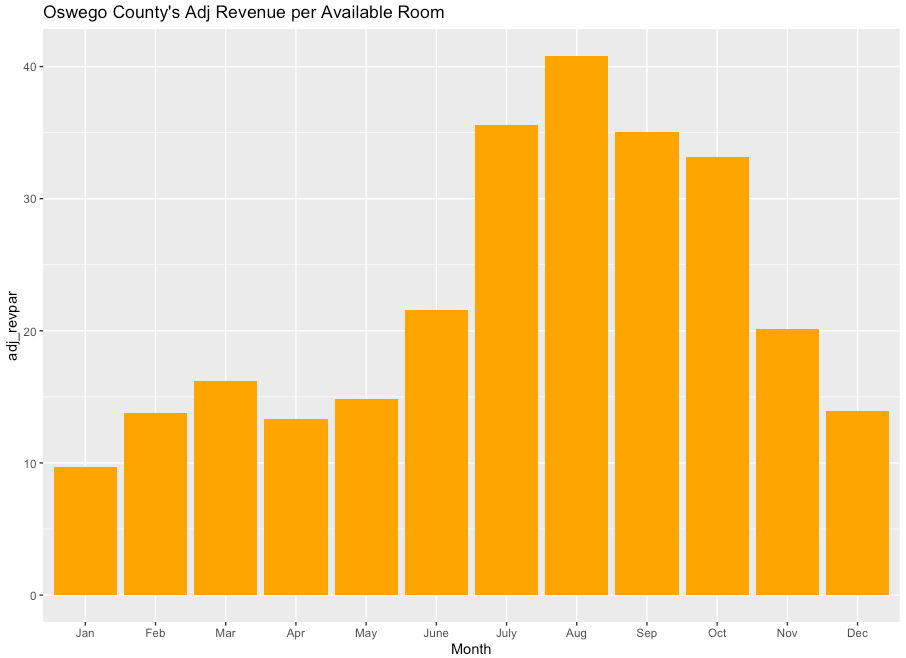
Figure 31

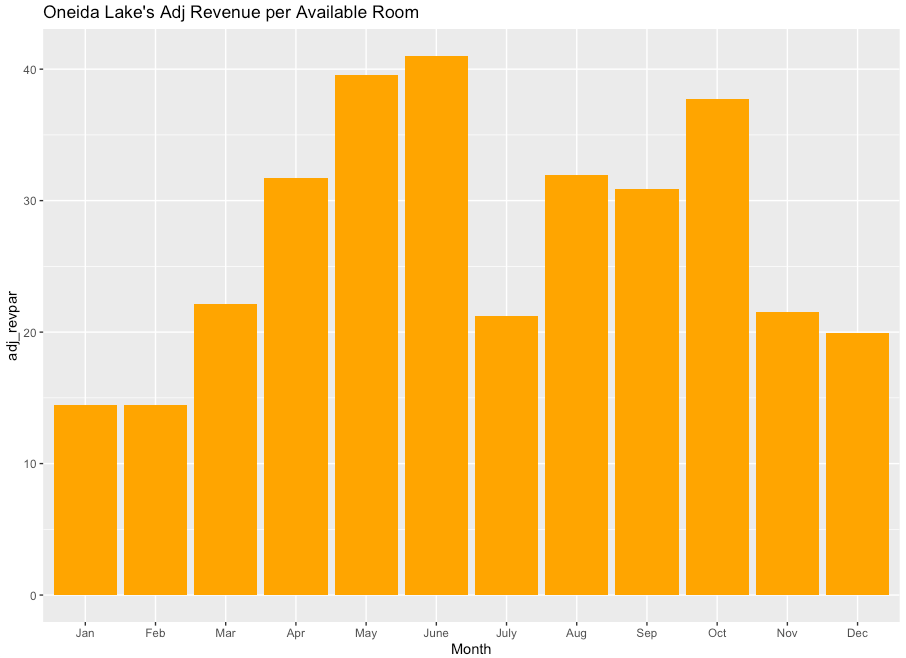
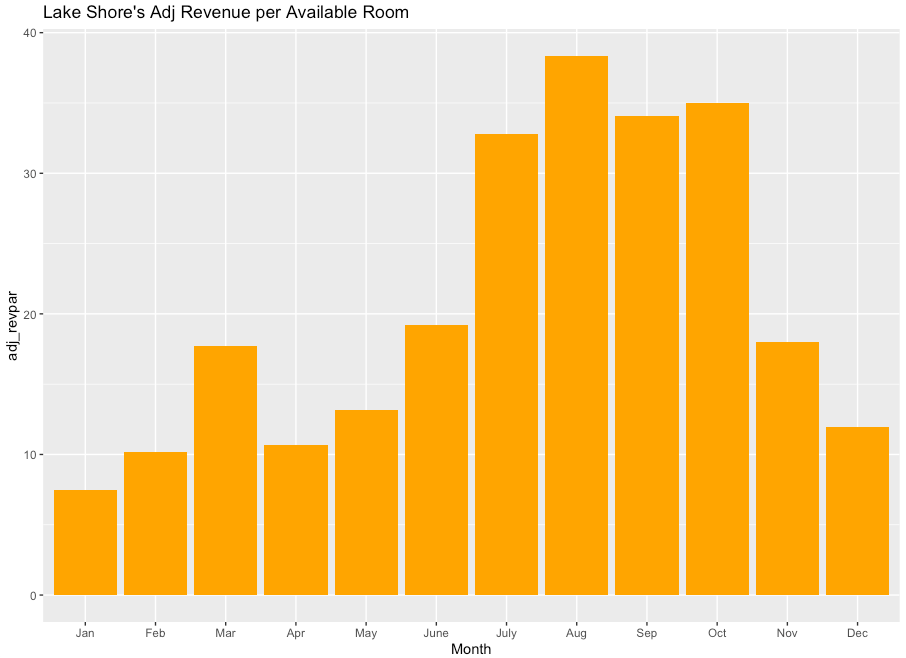
Figure 32

Figure 33

Figure 34

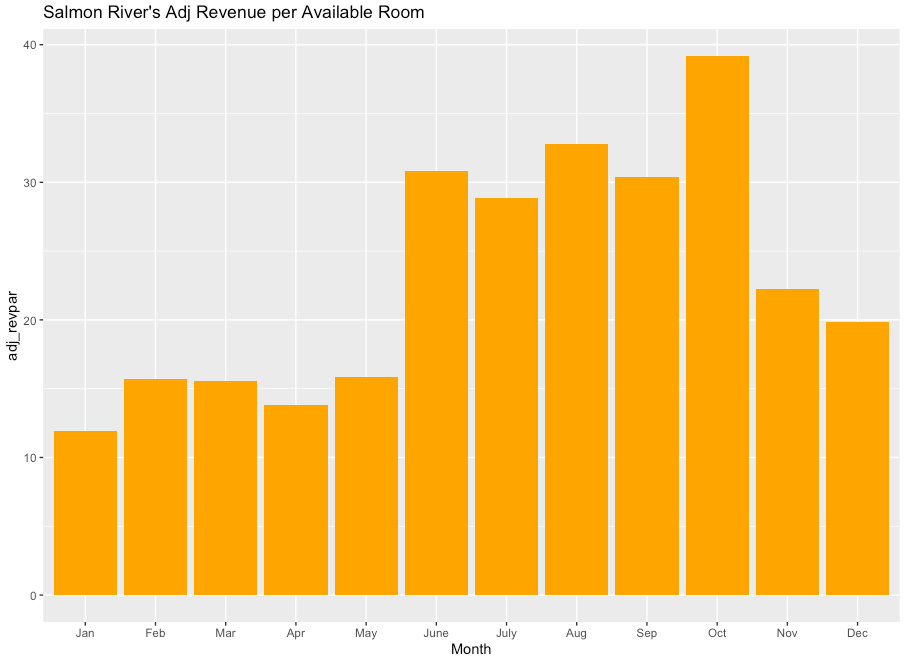


Figure 35

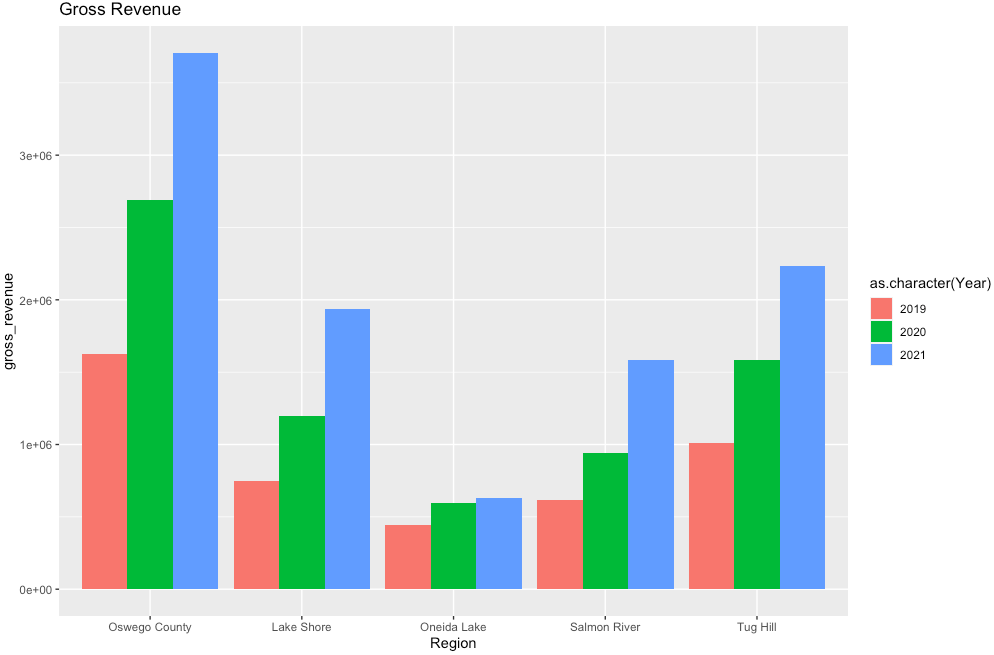
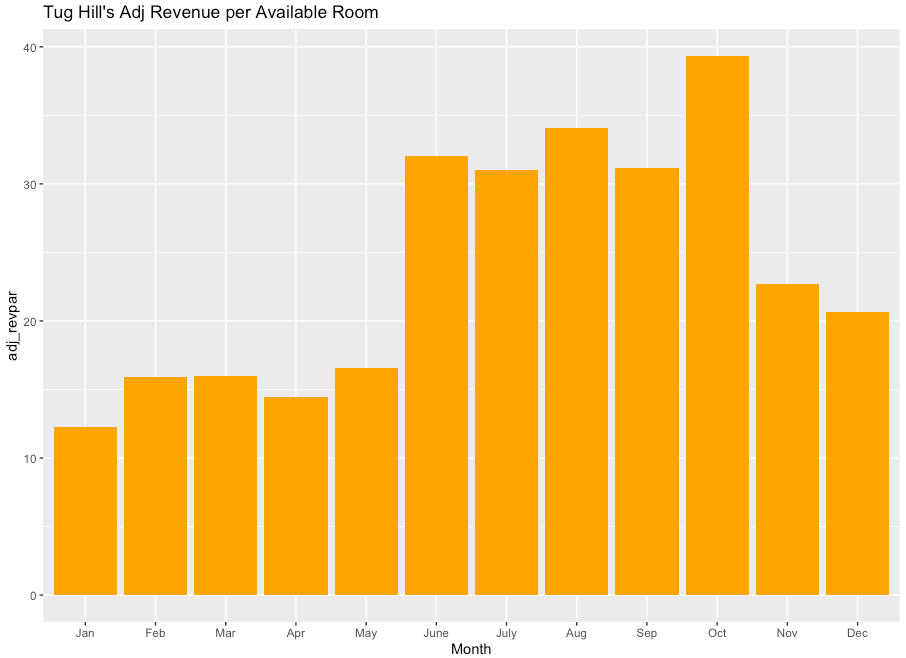
Figure 36

Figure 37

Figure 38