Assignment 1 Narrative
Name: Berke Derin Berktay
Github Username: brdrbr

Before we do anything else, first we need to be able to load the data and then use functions like head to observe the data and its columns. Which is what we do. Then, before we proceed with our ideas, first it is a good idea to perform some data wrangling to ensure that our data is suitable for all operations, is top-notch quality, only has the required variables present in order to prevent complexity, and has the required variables that are not currently present in order for us to implement our ideas further down the line. Therefore, first we start off by checking the amount of null and zero values for each column. In order to obtain these counts, we created a function called emptyandzero, which takes in the database. We do this using a function because down the line in our careers and education, for other projects we also need to perform this task and it is better to have a standard function that performs this to ensure we prevent repeatability. This was talked about in class, where Dr. Guha mentioned to ensure that we repeat as little code as possible by having a library of such methods that perform basic but routine operations like displaying the amounts of zeroe and null values for each column in a database. After creating the function, we call it and observe the results. We see that there are 35 empty values for the PROGRAM_NAME column, which does not matter in our case since we will delete this column later on since all of the important information in the PROGRAM_NAME column that is relevant is already present and contained in the SECTOR and ORGANIZATION_NAME column. Here, we also note that the amount of we the total length of the data, which is 50944, is equal to null bed entries (18545) + null room entries (32399). Therefore each row indeed has an entry regarding either beds or the rooms, which is how it should be. Then, we display the two entries where the PROGRAM_MODEL is empty. It turns out that these two cases are also the ones where the OVERNIGHT_SERVICE_TYPE and the PROGRAM_TYPE is also empty. Though a lot of useful informaiton like the program model is not available, I keep these two instead of deleting in this case since column entries like Sector and the service user count are not null, these two entries might still be useful later on.

Now, before getting started, we also need to perform some feature engineering. By looking at two random cases with the ID's 16051 and 16211 we see that for some ID's the day entries are way more than others. This is an important observation that we should keep in mind. Then, because of the forementioned reasons, I drop the PROGRAM_NAME column. Now, in order to use in the t-test studies further on, I need to have a very reliable continuous variable, therefore it is a good idea to create a column called OCCUPANCY_RATE, which is a percentage crafted using either the OCCUPIED_BEDS or the OCCUPIED_ROOMS column with the corresponding capacity columns. The value always ends up being in between 0 and 100, which makes sense. Next, I create a SEASON trend to observe possible trends for each season and compare them later on. Due to forementioned repeatability principle from the lecture, I optimize this by having a function called get_season do the job by taking in a date from the corresponding cell of an entry.

From now on, our data is ready and we need do some thinking to determine where to go from here. The most important thing to do is to determine which categorical columns to inspect. Here, I decided to use the idea of specificity. More specifically, the more specific divisions (more unique categories for the column) means the better the takeaways, the better and more specific focus that the government can have for the development purposes of the plans/models. We

want to observe the categorical variables in the scope of the previously crafted OCCUPANCY_RATE. We could conduct analysis on every metric, but the truth is some optimal categorical values for certain categories will contradict with some other optimal values for certain categories and therefore establishing a place with both optimal traits would be impossible. Therefore, by focusing on the caregorical variables with more possible categories, the government can have a more specific focus during planning, which can yield better results. The columns sector, overnight service type, program area, the program model, and the capacity are more concerned with the type of model to establish, so we can choose the most specific ones out of these (the ones with the most columns) in order to have a more specific agenda when coming up with building the types of models in the future. Like state before, it is better to simply prioritize the specific attributes since when building new models in the future, when determining their types, we can not always choose the best type of group in each category for the forementioned columns anyways since some might contradict each other's existence. Season is more about the timing, so it should be individually looked at to obtain takeaway ideas Because we want to focus on for more specific takeaways regarding the groupings, we look at: sector, overnight service type, and seasons vs only the factor of Occupation Rate and not service_user_count. I decided to not use service user count but only use the occupation rate since the service user count metric can be quite decieving since it completely disregards the capacity. For instance, a case with a huge number of user count, but if the occupancy rate of the case is around 30% for instance, it is not performing to its full capacity despite the huge number since there are a lot of rooms/beds remaining that are simply not used.

Now let us move on and observe the relation between the sector and the occupancy rate. First I wanted to craft some boxplots and violinplots to obtain some takeaways, Therefore, I allocated values for each category into new variables, and then created the two graphs using them. Here, I closely followed the procedure shown in class. The two graphs can be seen in figures 1 and 2. Then, I computed some crucial EDA's such as mean, max, IQR, and more for each type of category using a function called summary_stats. Again, a function to prevent repetition later on. The graph suggest that all types of SECTORs seem to have a relatively high occupancy rate, which clearly suggest that the capacity is not enough for the homeless in Toronto, and therefore further developments have to be made by the government. By looking at both graphs, we can see that the youth cases are relatively a bit more vacant (though still quite full) when compared to the other cases. This suggest that the government should focus on the development for the other types of sectors more since the occupancy rate overall seem to be fuller in the other cases. So the youth sector should not be the priority. It is important to note that a boxplot was not an ideal representation and therefore I also crafted a violin plot. This is because there is a max cap for the value of the occupancy rate, which is 100. These takeaways from the graph are also supported by the EDA, which states the YOUTH type as having the lowest mean overall. Though this does not mean that the youth are all able to find overnight places since the 75th percentile even for the youth is 100. Therefore, in at least 25% of the days the capacity for the youth is basically full, presumably during the winter months, when shelter is needed the most. The IQR is relatively higher also for the youth when compared to the others. This might be because since people their age are better equipped to work despite being homeless, they might be somehow getting by by sleeping in places like their work for instance. This is just an idea though, that can be tested later on as well.

Here, we can perform certain t-tests. It is the ideal scenario since we want to test if there is a significant difference between certain categories in terms of their occupancy rate results, which is a categorical variable. Though the data is clearly skewed for all due to the cap at 100 and lots of values accumulating around the 90-100 band, but because there are simply way too many entries for each case ~8000, the central limit theorem kicks in without an issue, but the CLT is not enough to change the fact that the variances of the groups are not equal due to their clearly unique skewnesses. Therefore, here what I did was that because the equal variance assumption does not hold, I compute a Welch's test between each categorical value combination and then transform the values using a popular log transformation technique to obtain the equal variance condition, and then compute a Student's 2 sample t-test between each categorical combination. Here, it is also important to note that normally, since we have 5 groups, I would have performed an ANOVA, which is also a type of t-test. But, since specifically simple t-tests were shown in class and emphasized in the assignment, I will perform a t-test for each pair unfortunately. Because we have 5 groups, we have 10 combinations. After performing the two types of t-tests, we observe the results. The results suggest that for literally all of the combinations and both types, the p-value is way too small from any critical value, (for our case we went with 0.05 by the way). This means that the difference between the means of the two groups are statistically significant. This suggests that all of the homeless groups are in different levels of having their requirements met, which means that for all of the groups there is a lot of work in the form of development and improvement that needs to be done. It is important to note that the student's t-test sometimes resulted in NaN values for the p-value and the t-statistic because of the way the data was log-transformed, but I could not fix this issue no matter how hard I tried. For the t-tests of both types, I have also used a function to prevent repetitiveness.

Now, after observing the relationship between the sector and the occupancy rate, as aforementioned, we also observed the relationship between the overnight service type and the occupancy rate. We followed the exact same procedure of forming a boxplot and a violinplot, conducting EDA for each type of category, and finally conducting the t-tests of two types after performing the required transformation for the Student's t-test. Note that I also use two functions to perform the tests since there were way too many combinations here due to the 7 types of categories existing. Here, the results and the takeaways are quite interesting to say the least. The non-NaN p-values all again suggest significant differences between the groups. Again, this means every site needs to be developed more. But more interestingly, the graphs, shown in figures 3 and 4, suggest that the isolation/recovery sites are significantly empty overall. However, this might be because of the policy of social distancing since the purpose of these sites is to basically deal with the covid pandemic. Therefore, it is somewhat expected for them to be empty. But an idea to pursue from this key-takeaway is to maybe have these empty isolation/recovery sites partially function as other types of sites such as warming centres since the capacity for the already existing sites of other types is clearly not enough due to the average capacity being around 90's again. By ensuring a clean environment, the disease would not spread further and more homeless people can also benefit from these sites. This is a great idea in my opinion. It is also worthy to note that, though not to the same extent, the 24-hour women drop-in sites are relatively emptier than the other types of sites. We can transfer people that could not get into the interim housing sites for instance to these places when there are

vacancies since the interim housing sites are completely full almost 100% of the time, as it can be seen from the plots. These ideas can be easily implemented and benefit the homeless community.

Lastly, let us observe the seasons. As predicted, the bar graph labelled as Figure 5 suggest that the occupancy rate is lower in the summer months than the winter months. Interestingly, the occupancy rate in the fall is significantly higher than it is in the spring. Which suggest more temporary projects to be applied in the fall season than in the spring season. I also looked at the occupancy rate based on type during the seasons and graphed it, as seen on Figure 6 to determine if certain types of shelters are consistently over capacity or have higher occupancy rates during specific seasons. The graph suggest that the 24-hour women dropin is significantly emptier during the spring and summer, which suggests that these dropins can be used for different purposes partially, especially during these months, and that the isolation/recovery sites are significantly emptier during the summer and autumn, which suggests that the spread of covid is more rapid during the winter and spring.

APPENDIX:

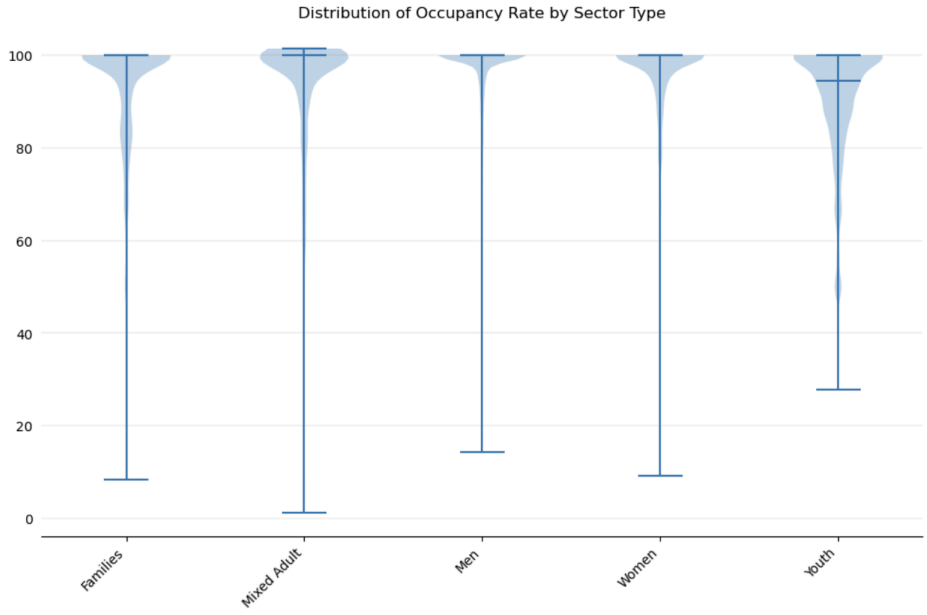

Figure 1

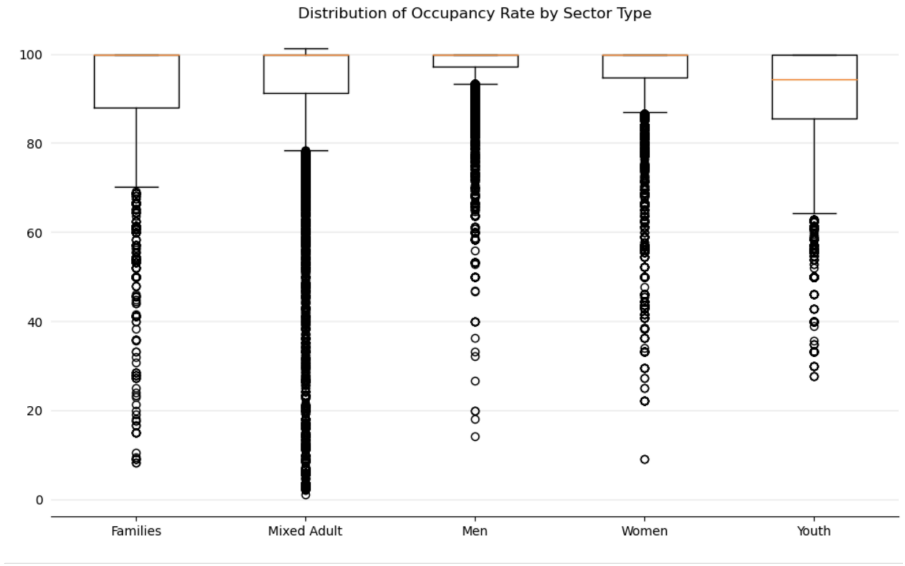

Figure 2

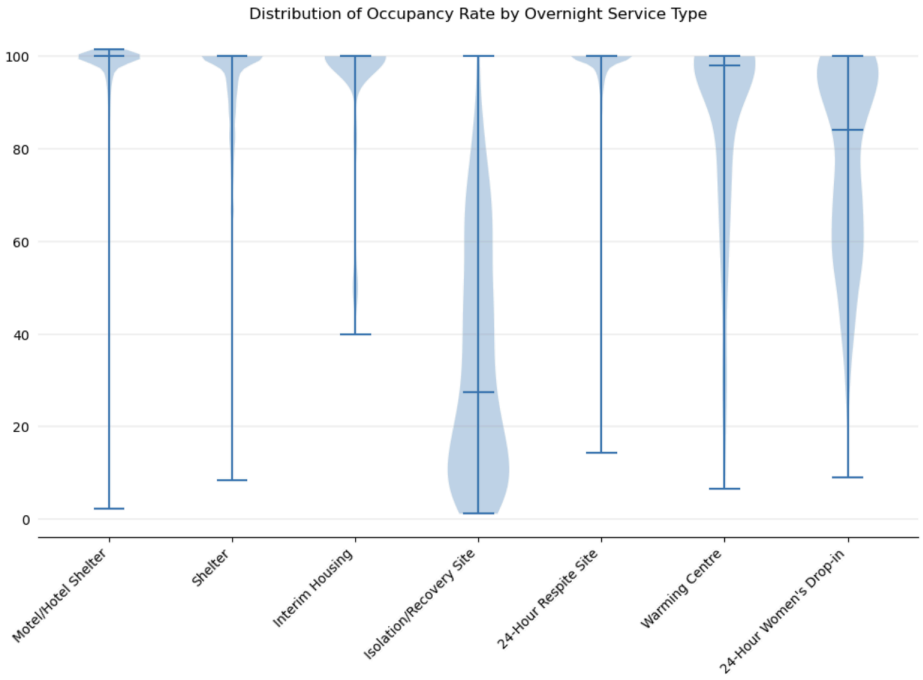

Figure 3

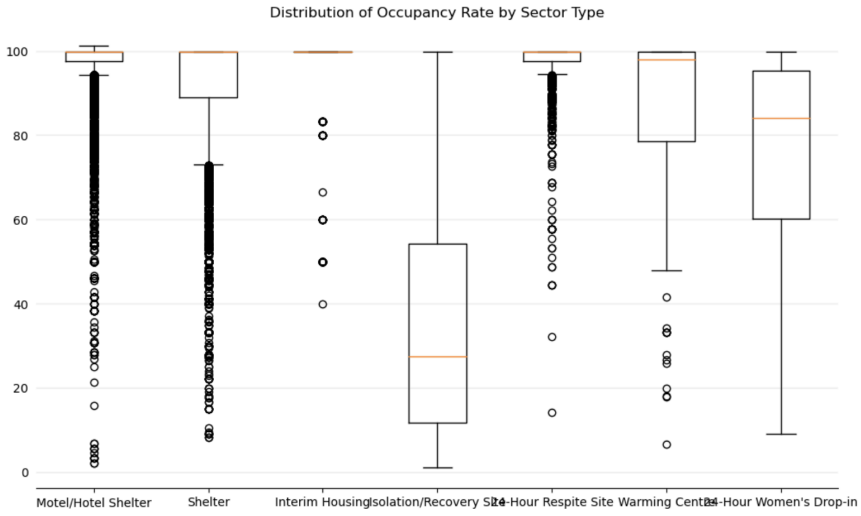

Figure 4

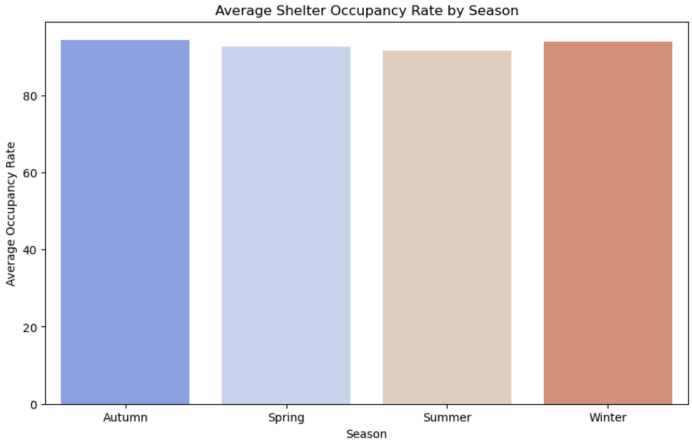

Figure 5

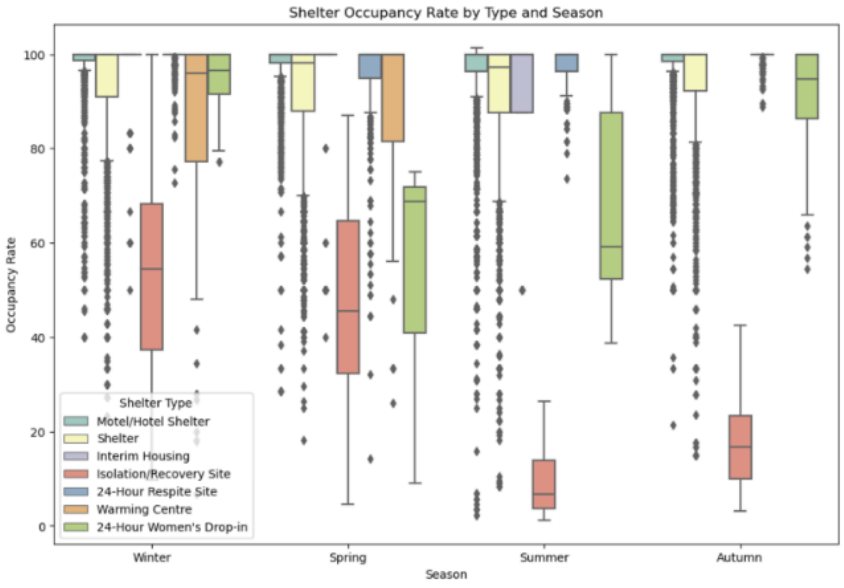

Figure 6