Task 3: Documentation of Cyclability Analysis

1. Dataset Description

Data sources for the initial 5 datasets are from canvas. The two (2) additional datasets were obtained from reliable online sources which would be further elaborated below. The source that we obtained the additional datasets were reliable and the datasets did not have many missing or 'incorrect' looking values.

For the original five (5) datasets that were provided on Canvas, there was no metadata provided to aid in understanding the data and no legal guidelines as to how we might use the data. Due to this, we filtered and removed the data in a manner we felt was appropriate. For the usability of the additional two (2) datasets, we ensured that that the way that we used it did not infringe on any copyrights and we ensured we used it within the scope permitted by the Australian Bureau of Statistics.

Our dataset Number of Vehicles was obtained from an original dataset - *Census Time Series 2016, 2011, 2006: T22 Number of Motor Vehicles by Dwelling records (SA2+)* sourced from the Australian Bureau of Statistics. Our data that we used for our Number of Vehicles dataset was obtained by extracting the relevant data. It includes the area ids, the year that the data was recorded, and the number of vehicles in each area.
http://stat.data.abs.gov.au/Index.aspx?DataSetCode=ABS_C16_T22_TS_SA

The Spatial shape file came from the Australian Bureau of Statistics - *Statistical Area Level 2 (SA2) ASGS Ed 2016 Digital Boundaries in ESRI Shapefile Format*.
https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument&fbclid=IwAR19QCKA_AAbSWofoPV8prxRWqNoUHW63lwANUmi_3jfjAGGUwpVRF_WBtk#Data

Pre-processing steps

1) We loaded the original 5 datasets and an additional 2 (One spatial file and one csv file) into Jupyter notebook and read them using python's imported library 'pandas'. We labelled each dataset as follows;

df (Neighbourhood Data)
df1 (Bike Sharing Pods)
df2 (Business Stats)
df3 (Census Stats)
df4 (Statistical Areas)
df5 (Number of Vehicles)
sf (Shape file)

2) We checked the 'cleanliness' of the datasets by using *dataframe.isnull().sum()*. This listed out each column of the data frames and counted and provided the number of 'NaN' values each column had. If all values were 0, we deemed it a 'cleaned' dataset. After performing the check, we discovered that **'Neighbourhood data'**, **'Business Statistics'** and **'Census Data'** had 'NaN' values indicating that 'cleaning' would be required before being able to proceed.

3) As there was no metadata provided with these datasets, we had to make assumptions when cleaning the datasets. We decided to 'clean' the 3 'NaN' datasets based on careful and thorough interpretations. For the **Census statistic** dataset, we used *dataframe = dataframe.dropna()* to drop the entire rows that contained 'NaN' values, with the understanding that no households have null income. For **Neighbourhood** and **BusinessStats** datasets, we replaced the 'NaN' values with 0s. These 0s here have valuable meanings, that in those areas with 0s, there was no available business of that type in the area and no inhabitants in that area.

4) The cleaning process for the **Number of Vehicles** was slightly different from others. After the original processing to ensure there were no NaN or 'incorrect' values, we removed 4 columns were not wanted by using *dataframe = dataframe.drop(columns="column_name")*. After which, we renamed the remaining columns for consistency by *dataframe.rename(columns={'original_name':'desired_name'}, inplace=True)*

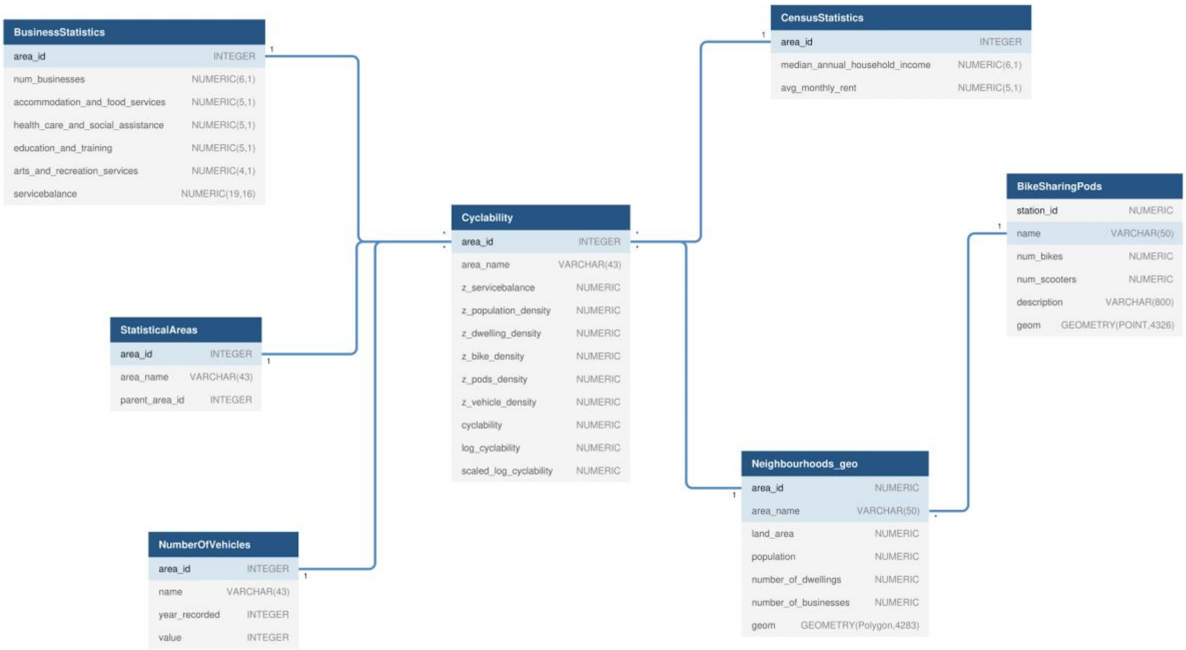
5) For our shape file which contained the suitable geometry and information required for us to perform the spatial join, there was no cleaning required for this dataset. The only step we had to take was to obtain the geometry values for 'Greater Sydney' only, and combined it with the Neighbourhood dataset. Then we uploaded the new Neighbourhood dataset to SQL database with name **'Neighbourhoods_geo'.**

6) For the final stage of cleaning, we ensured that each column had the appropriate data-type. For example, if the column was supposed to contain a String type, but contained a numeric value instead, we would not include it and filter it out as inconsistent values.

7) Finally, we outputted the cleaned dataset as .csv files for future use.

2. Database Description

The diagram below, presents all our created tables that are stored within the ‘Public Schema’. For all the data frames, we made a table and schema with the ‘Primary Key’ being ‘area_id’ except for BikeSharingPods whose ‘Primary Key’ is ‘station_id’.



An index is used to find all rows that match a column based on the query and locate all exact matches from the table data. For example, we created an index on ‘area_id’ that is a ‘Primary Key’, this means that when the query is passed the server would be able to locate the table data faster as it will not have to manually search through every table data row to see if they match the search query.

We created a ‘Spatial’ Index on geometry column of neighbourhoods_geo, named “neighbourhoods_geom_index”. It sped up the process of spatial join from 10 sec 607 msec to 8 sec 848 msec. We also created a 'Primary Key' index on ‘area_id’ of SJ table, named ‘area_id_index', which decreased the table join time from 400 msec to 271 msec.

A Foreign Key is part of a different table (Child Table) that has a different ‘Primary Key’. The Foreign Key of the Child Table is used to reference the ‘Primary Key’ of the Parent Table(s). This links the Parent Table(s) and Child Table(s) without disturbing the link between the Parent Table(s) and other tables with the same Primary Key.

We did not see the need to create a Foreign Key for our Database as none of our Tables required the use of it.

3. Cyclability Analysis

Several adjustments were made to the formula in computing the cyclability scores, formula shown below:

$$\boldsymbol{Cyclability} = \pm \boldsymbol{log}(\left|-\boldsymbol{Z}_{\quad \boldsymbol{population\ density})} + \boldsymbol{Z}_{\boldsymbol{dwelling\ density}} + \boldsymbol{Z}_{\boldsymbol{service\ balance}} + \boldsymbol{Z}_{\boldsymbol{bike\ density}} + \boldsymbol{Z}_{\boldsymbol{pods\ density}} - \boldsymbol{Z}_{\boldsymbol{vehicle\ density}}\right|)$$

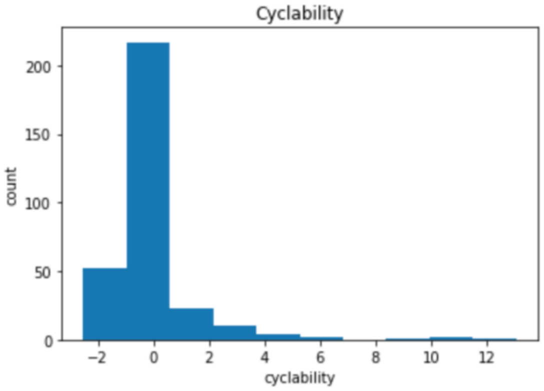
[If the sum of all z scores are negative, we put a negative sign in front of the log operator]

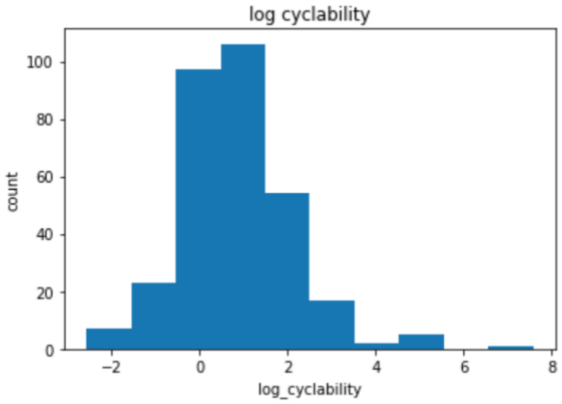
1) Instead of adding all z-scores, we subtracted the population and the vehicles density z-scores since they have negative impacts on the cyclability score (reduces the cyclability of an area of high population and vehicles density).

High population density indicates more people living in the region. A crowded areas are less ideal for biking, since bikers have to be more cautious about the safety of the pedestrians that places dangers on both bikers and pedestrians. Similarly, vehicle density is interpreted as the number of vehicles on the road by area, when the density is high, there are more cars on the road and increases the likelihood of car accidents between drivers and bikers.

Z scores measures how many standard deviations the values are away from the mean, and the final cyclability score is a sum of a number of z scores. A high z score for each field should suggest a beneficial impact on cyclability, since both population density and vehicle density have negative effects on the cyclability scores, we subtracted them in computing the cyclability score to reverse the impacts and keep consistent to the interpretation of cyclability score.

2) Logarithmic return is computed on the cyclability score to obtain log_cycability.
Histogram for cyclability (left) to visualise the spread of the data. And it can be clearly identified that the cyclability variable is highly right or positively skewed, with a high skewness of 4.337 and kurtosis of 23.871, shown by the fat tail that indicates with extreme outliers. Hence we computed the logarithmic return of the variable cyclability, the log return normalises the data outputting a normal distribution (as shown on the right histogram). Also the influences of the extreme outliers are removed, as logarithmic function is not affected by the outliers effects. When computing the logarithmic values of the cyclability score we take log of the positive values whilst take log of the absolute value of the negative values then assign it with negative signs.



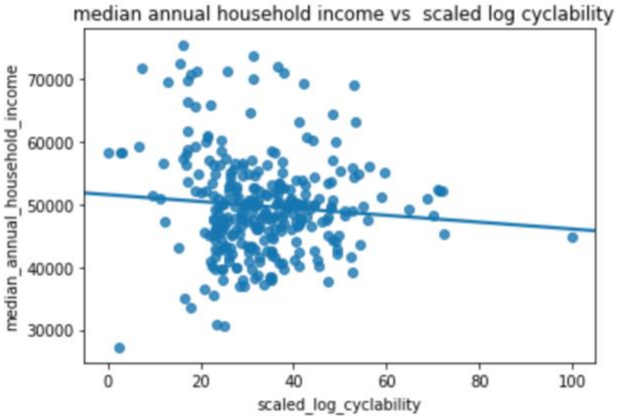


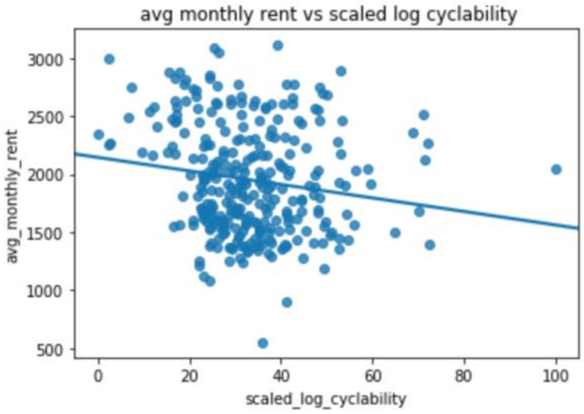
3) Then for easier interpretation of the cyclability score, we scaled the **log_cyclability** into a range of 0 to 100. The following formula is utilised to compute the cyclability score out of 100. Thus we obtain variable scaled_log_cyclability.
The formula is founded from the following website:
https://stats.stackexchange.com/questions/281162/scale-a-number-between-a-range

$$score_{\quad normalised} = (100 - 0)\,\frac{score - min(score)}{max(score) - min(score)} + 0$$

4. Correlation Analysis

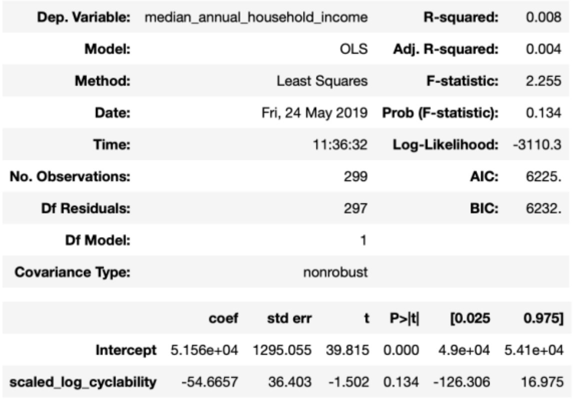
The correlation values are computed in python, values obtained are -0.0868 for **median household income vs cyclability** and -0.1585 for **average monthly rent vs cyclability.** Both pairs have a negative correlation and a stronger correlation is clear for the latter pair. Further analysis on the potential relationships are through importing the seaborn function in python to construct scatter plots for visualisation. (Left **median annual household income vs scaled cyclability,** right **average monthly rent vs scaled cyclability**):





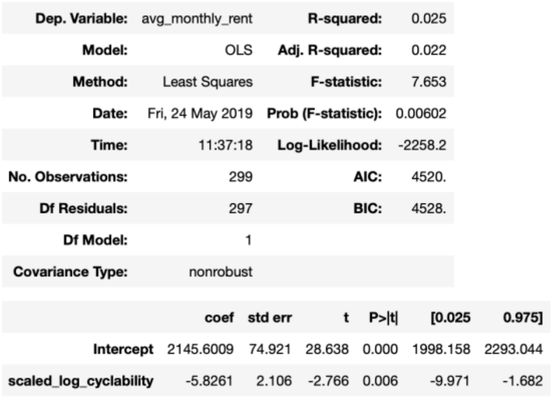
The regression line along with the scatter plots allows us to observe that the median household income in the given suburbs have a clear negative linear relationship with the cyclability scores. Similarly, a stronger negative linear relationship (greater negative slope) exists between the cyclability score and average monthly rent.

For further interpretation of the correlations, we imported the statsmodels.formula.api to compute the linear regression model and determine the strengths of these relationships. (Left **median annual household income vs scaled cyclability**, right **average monthly rent vs scaled cyclability**):



Dep. Variable:	median_annual_household_income	**R-squared:**	0.008
Model:	OLS	**Adj. R-squared:**	0.004
Method:	Least Squares	**F-statistic:**	2.255
Date:	Fri, 24 May 2019	**Prob (F-statistic):**	0.134
Time:	11:36:32	**Log-Likelihood:**	-3110.3
No. Observations:	299	**AIC:**	6225.
Df Residuals:	297	**BIC:**	6232.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5.156e+04	1295.055	39.815	0.000	4.9e+04	5.41e+04
scaled_log_cyclability	-54.6657	36.403	-1.502	0.134	-126.306	16.975



Dep. Variable:	avg_monthly_rent	**R-squared:**	0.025
Model:	OLS	**Adj. R-squared:**	0.022
Method:	Least Squares	**F-statistic:**	7.653
Date:	Fri, 24 May 2019	**Prob (F-statistic):**	0.00602
Time:	11:37:18	**Log-Likelihood:**	-2258.2
No. Observations:	299	**AIC:**	4520.
Df Residuals:	297	**BIC:**	4528.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2145.6009	74.921	28.638	0.000	1998.158	2293.044
scaled_log_cyclability	-5.8261	2.106	-2.766	0.006	-9.971	-1.682

From the outputs above we can derive the linear prediction equation for **median household income vs cyclability:**

$$\text{median\_}\widehat{\text{houshold}}\text{\_income} = 51562.198057 - 54.665659 \times \text{scaled\_cyclability}$$

Simple interpretation of this prediction model: an increase of 1 in the scaled_cyclability score we will expect a decrease of $54.665659 in the median_household_income on average, based on our sample data in Sydney. This is clear to us that an area (suburb) with a high cyclability score have a lower median household income. Hence, a suburb with lower cyclability score will have a higher median household income (e.g. when the cyclability score is 0 we have an estimated median household income of $51562.198057)

This correlation aligns with our hypothesis, as a more cyclable area is often less dense in population,vehicles , and dwellings and are likely to be in the outer region of Sydney rather than in the city. Therefore, this impacts the median household income to be low, due to less opportunity and job limitations in the countryside.

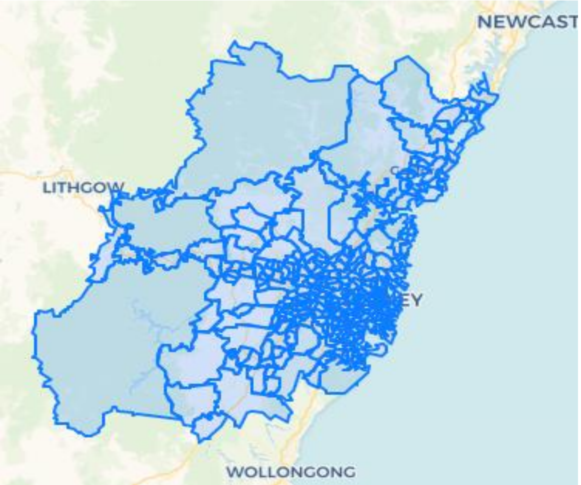
From the value of R^2, we can also determine the strength of this relationship, the model has a $R^2$of 0.008, which indicates the independent variable can explain the dependent variable at 0.8% of the dataset. This is a weak negative relationship of the median household income and the scaled cyclability scores.

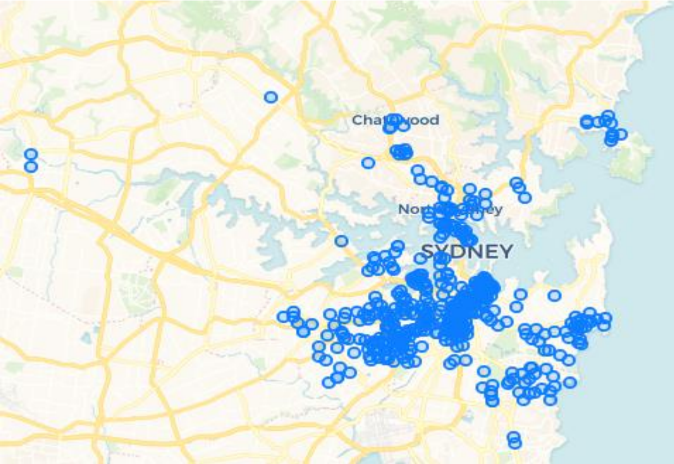
Similarly, the prediction linear model for average monthly rent is:

$$\widehat{\text{avg\_monthly\_rent}} = 2145.600893 - 5.826126 \times \text{scaled\_cyclability}$$

An increase in 1 of the scaled cyclability score, there will be a decrease of $5.826126 in the average monthly rent. Similar to the previous relationship, a higher cyclability score will expect a lower average monthly rent in the neighbourhoods. However, compared to the previous pair it has a less negative relationship (increasing by 1 only reduces by $5.826126). Areas with lower cyclability scores have relatively higher average monthly rent (e.g. area with 0 cyclability score has an estimation of $2145.600893 average monthly rent.) This indicates that area that are more cyclable have lower average monthly rent, and this is likely due to the areas of higher cyclability score is further away from the cities and areas of popular residential areas, thus has a lower mean monthly rent.
Interpreting the R^2 value which is 0.025, indicates 2.5% of the variables are explained, a stronger correlation between the **average monthly rent vs scaled cyclability** compared to **median household income vs scaled cyclability.**

5. Data visualisation





Error analysis

There are some points missing after the spatial join, this is because the neighbourhoods geometry data we collected from ABS has 13 invalid geometries. Hence the spatial join with these polygons failed, and some points were not appeared in the spatial joined table. This might affect the accuracy of bike_density and pods_density, and therefore the accuracy of cyclability score is not guaranteed.