Finally, there is a

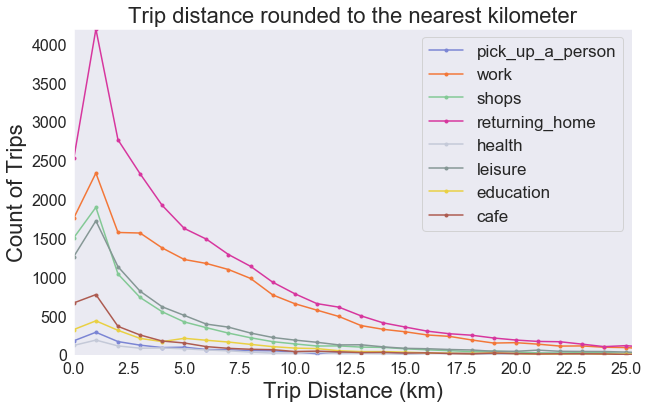
Data collection [for survey data] driven by data availability or convenience of data collection rather than by domain knowledge, theory, or insight into the process(es) of interest. (An *et al.*, 2015).

After outlier removal

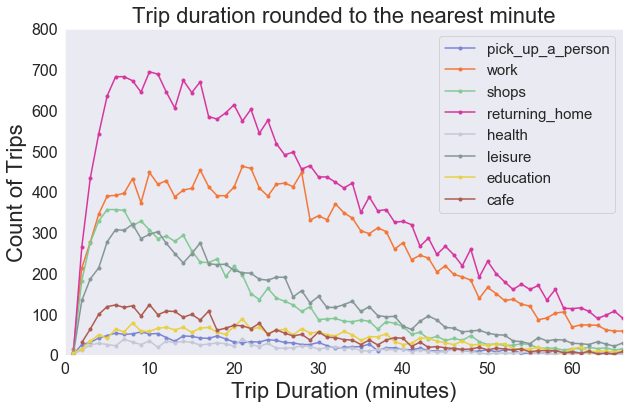
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **std** | **min** | **25%** | **Median** | **75%** | **max** | **N** |
| **Distance (m)** | 6634 | 9927 | 50 | 840 | 3147 | 8092 | 99810 | 177938 |
| **Duration (sec)** | 1537 | 1285 | 60 | 616 | 1204 | 2081 | 10799 | 177938 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wolf *et al.* (2001) | Distance; duration; land-use matching | Georgia, USA 3 days  ﻿March 23 through April 28, 2000 | 13 (156 trips) | 93% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probabilistic Methods: Multinomial Logit Model | | | | |
| Oliveria *et al.* (2014) | Duration; Mode; Land use; Personal locations proximity | Georgia, USA 2011 | 1,354 (﻿10,512 trips) | 70% (95-97% for work and school) |

****

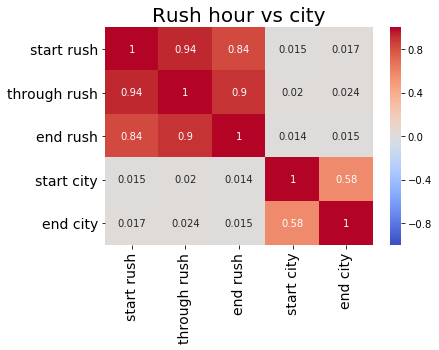
**Figure 4.X**

****

**Figure 4.X**

Overall, the rush hour and city labels are found to exhibit statistically significantly (*p-value<0.05*) positive correlation across the trips, as determined by a spearman’s rank correlation coefficient (ρ) (**Figure 4.8**). Notably, there is no discernable correlation

we see weaker positive correlation between trips that begin and end in the city.



**Figure 4.X** Correlation matrix (statistical significance below a 0.05 confidence interval indicated by white text; *n=73,029*)

KS test vs norm for grouped 1 hour

cafe KstestResult(statistic=0.5551535507713675, pvalue=3.287213099129875e-205)

education KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

health KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

leisure KstestResult(statistic=0.7251252221594695, pvalue=0.0)

not\_available KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

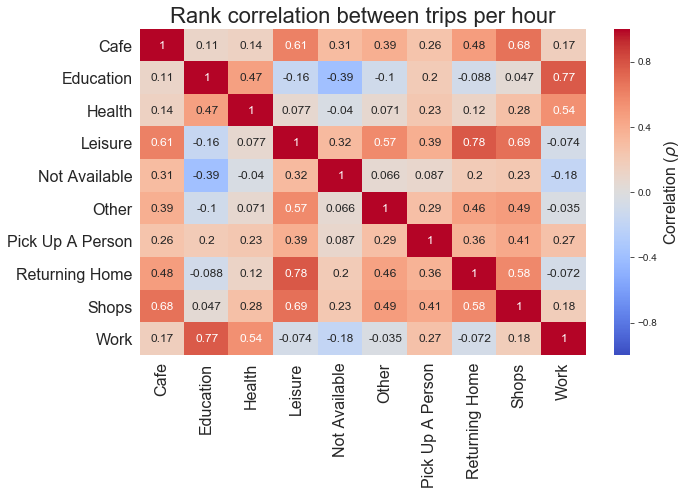
other KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

pick\_up\_a\_person KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

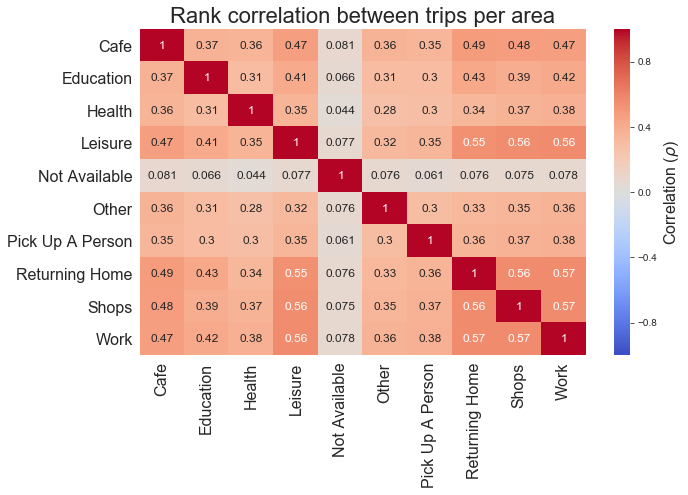
returning\_home KstestResult(statistic=0.8370232391566367, pvalue=0.0)

shops KstestResult(statistic=0.6061450521877981, pvalue=3.586226223248267e-249)

work KstestResult(statistic=0.749204542272784, pvalue=0.0)

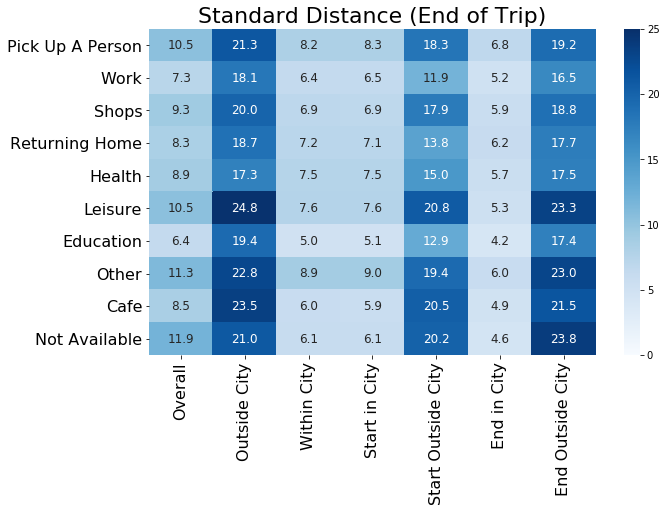
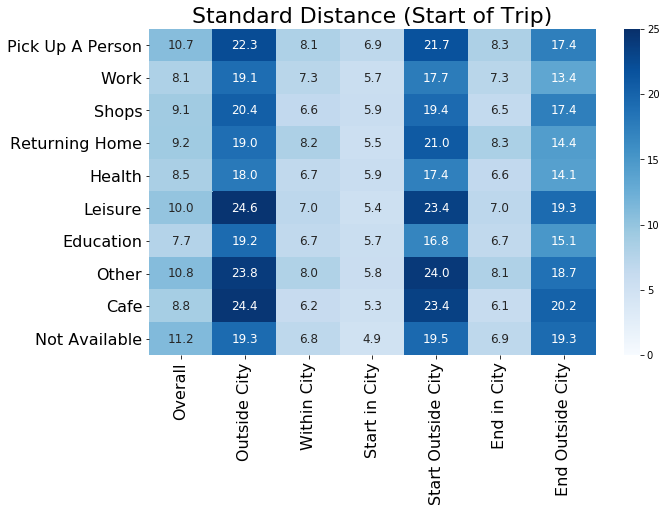


**Figure 4.X**



**Figure 4.X**

Standard distances between each unique trip purpose and



**Figure 4.17** Heatmap plots showing the standard distance between start and end points of the trips grouped by individual trip purposes.

For a measure of spatial dispersion within the model inputs, the Standard Deviational Distance is calculated using *Pysal* on all the start and end points of the trips. This statistical measure gives us an indication of how far trips for given purposes are spread across Montreal (Rey & Kang, 2019). Finally, to determine the degree of spatial correlation between the trip purpose classes we use a spearman’s rank correlation matrix to compare the count of each purpose in each of the 7,046 DAs.

* This non-parametric correlation co-efficient is preferred after initial Kolmogrov-Smirnov test suggest the

New Purpose labels:

Based on correlation

"1":["shops","leisure","cafe","returning\_home"],"2":["education"],\

"3":["health"],"4":["pick\_up\_a\_person"],"5":["work"]}

1 47464

5 18950

2 2769

4 1574

3 1044

Based on discovery of trip purposes that share spatial and temporal dependencies we further subset the 10 unique trip purpose classes into . We also opt to remove trip classified as ‘Other’ and ‘Not Available’. It is hoped that these will improve the performance of the classifiers (Jahromi *et al*., 2016).

Table of New Groups:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Label | Purpose Types |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |

Extra:

The improvement of mobile networks and the geolocation abilities of smartphones in recent years, means we currently have an opportunity to study cities through data resulting from mobile GPS traces (Li *et al.*, 2016; Patterson *et al.*, 2019).

Mobile phones as sensors (ref)

Extra:

Respondents:

B & M: 1104

Oliv 1,354

Xiao 321

Mont 156

Kim 793

Yaz 6845

Zhu 10,372

Geographically weighted regression between weather (temperature and precipitation; 3.2.2) and mode/purpose in given areas. A model will be built for each individual mode and purpose. This form of regression makes use of the queen’s case contiguity matrix to account for the spatial lag. It is hoped that this will give an idea of sensitivity of transport in certain regions to weather i.e. Downtown will likely not be as effected by rain.

### 3.4.3 Spatio-Temporal Investigation of Model Inputs

- directionality throughout day

- space-time calendar (after Arribas-Bel & Tranos, 2017) for given regions of Montreal

- space-time interaction tests in pysal

- space-time evenness grid

- cross-correlation and Coefficient of determination between given regions/DAs of Montreal

“We aim to create conditions for ‘effective’ space-time forecasting” (Yue & Yeh, 2008).

ST-KDE:

See Wei *et al.* (2018) (Also uses grid-ified data)

To-Add:

* [On visualisation of big geodata] We should carefully generalize, e.g., emphasize the important while removing the unimportant, (Li *et al.*, 2016)
* Reason for: Insight into those spatial and temporal trends can improve the performance of Intelligent Transportation Systems (ITS).(Tayyab *et al.*, 2014)

Sentences:

The casual links have been harder to prove in the literature

Owing to the…

To Add:

* This has affected probabilistic models over machine learning models (Oliveria et al., 2015)

Arguably, there is a need for more encompassing approach to the modelling inputs has allowed more generic model creation that can applied to a range of different purposes (as we do not know which purposes rely on which i.e. POI may be important for tourist purpose but not for work purposes).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gong *et al.* (2018) | POI; Socio-demographics; Temporal Features; Trip Duration; Weather | ﻿Hakodate City, Japan 2012–2013 | (﻿9981 trips) | n/a |

Spatial-Temporal:

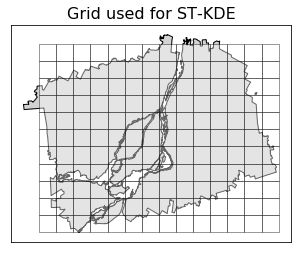
ST-K

Space-time Ripley’s K and spatial-temporal randomness -> end to start

[One STKDE Ripley’s K]

Space-time KDE end to start

ST-KDE



**Figure 4.X**

[Some sort of plot for KDE]

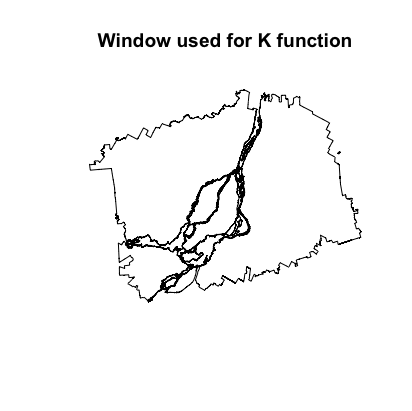
Join the ST-KDE back to the data -> if there a trip at a particular time and space (a grid region) can be flagged as in the given cluster (0,1)

11 columns of STKDE cluster

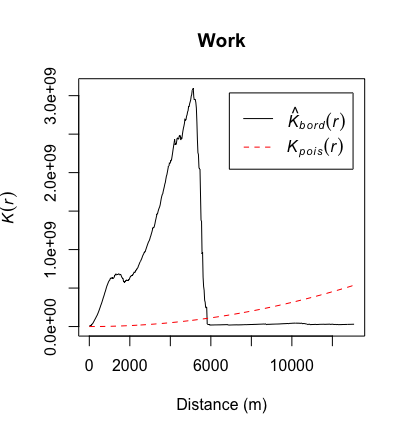
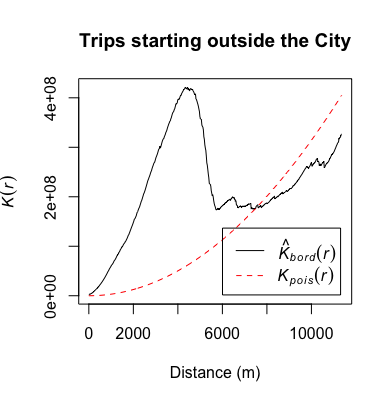
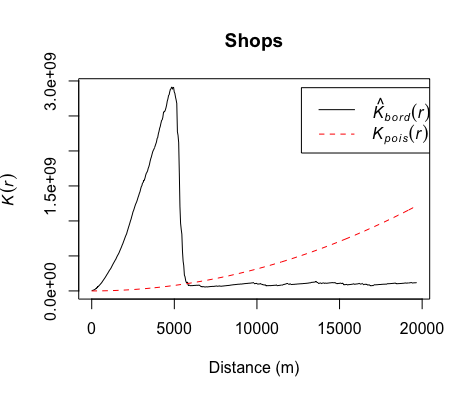
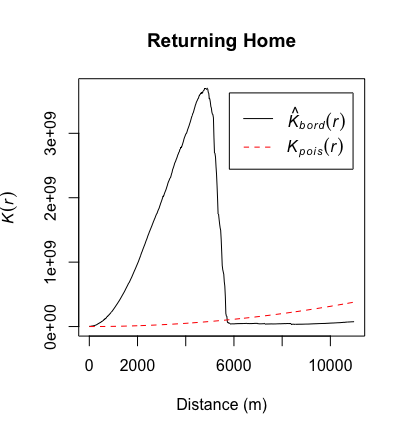
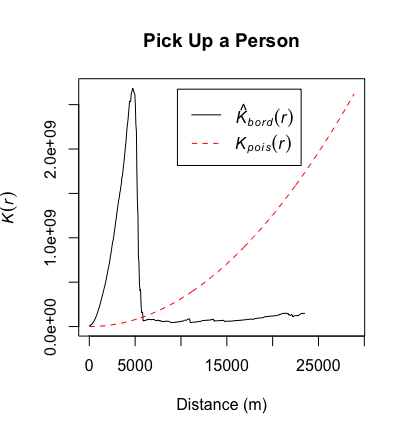
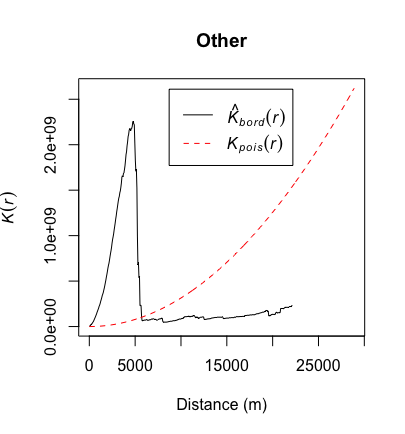
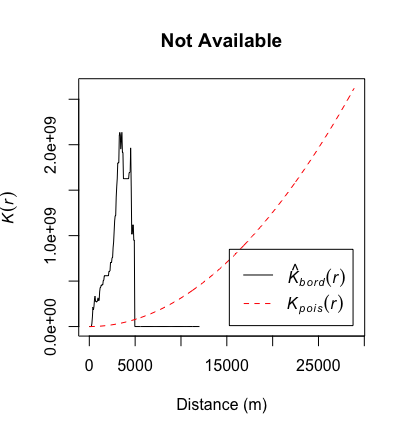
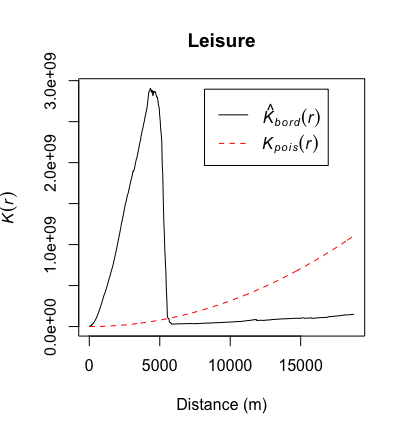
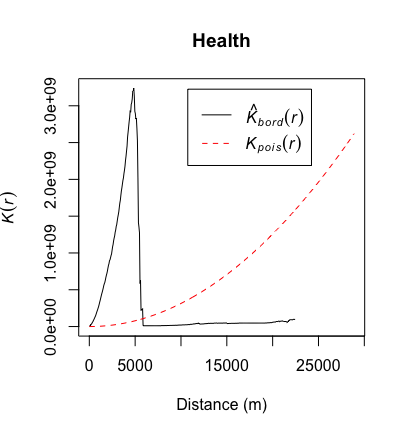
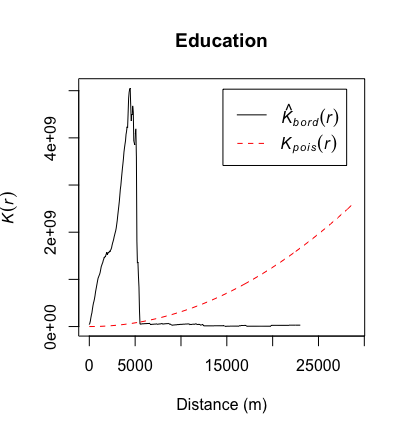
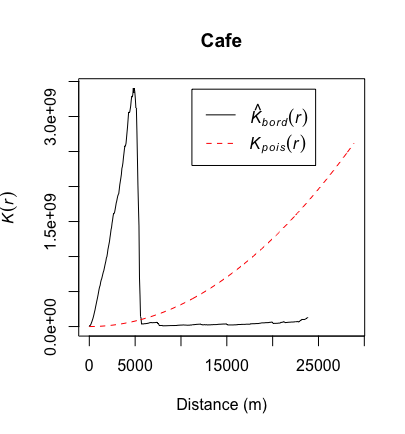
For a given trip, dissemination areas within the intersection are given a value of 1 (red) and all other areas are given a value of 0 (grey). For the spatial analysis in this report, data has been grouped using *geopandas groupby* function into individual modes and purposes of travel (see **Table 3.1** in section 3.2.1 for types) and then visualisations have been produced from each showing the total intersection count in each dissemination area across the entire study period (see X.X).

**Figure 3.7** Example of the spatial join between a route and the underlying dissemination areas (route in **blue**; overlapping dissemination areas in **red**)

Ripley’s K on start and end -> spatial randomness



**Figure 4.X** Window used for Ripley’s K function



“Using this window, a points pattern object is created from the Treasure Hunt location coordinates and estimations for Ripley’s K is calculated. This estimation of K is plotted in Figure 5 against a Poisson distribution (which indicates Complete Spatial Randomness of Treasure Hunt locations). As shown, the trend in estimated K is far higher than the Poisson distribtion implying there is clustering and spatial dependence of Treasure Hunt locations across London.”

We assess spatial dependency of trips from each type of trip purpose using Ripley’s K-function calculated using the R’s *spatstat* library. Specifically, the Ripley’s K can be used as a measure of spatial distribution of start and end points of the trips indicating whether there is Complete Spatial Randomness within the data (Dixon, 2002). These

Moreover, we detailed in **Table 3.6**. we divide the data into four distinct subsets from which all models are trained, The purpose of this is to evaluate the disparity in performance within the model when only looking at one broad spatial or temporal category.

**Table 3.6** Subsets used to evaluate classification model accuracy

|  |  |
| --- | --- |
| *Subset Name* | *Description* |
| City | Trips starting o ending in the city |
| Non-City | Trips starting and ending outside the city |
| Rush Hour | Trips starting or ending in the Rush Hour |
| Off-Peak | Trips starting and ending in Off-Peak |

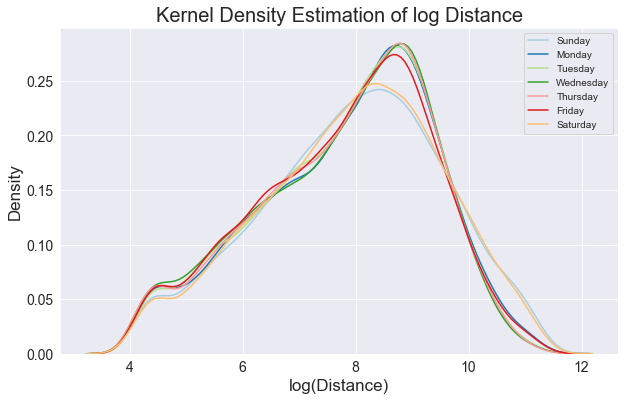
Individually, decision trees represent the probability of one of a set of outputs given an input condition, in RF classification models, decision trees are used in combination with (; find other ref)

Boostrap aggregation (or bagging)

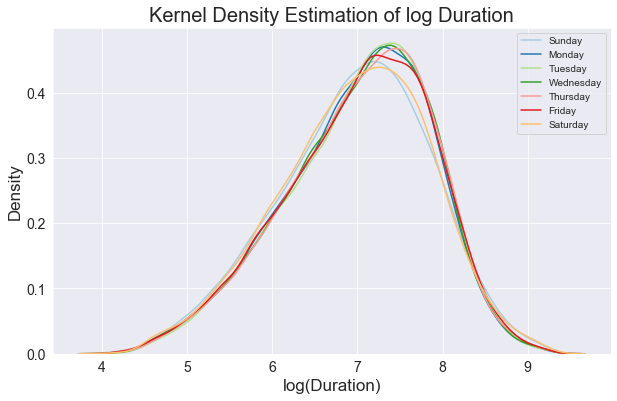
The data is hence, normalised for the purpose of the classification models using by taking the natural log. **Figure 4.3+4.4**

Log distance skew: -0.37300090670060865 kurtosis: -0.5147541361828902

Log duration skew: -0.38136680556957275 kurtosis: -0.21730048341457042



**Figure 4.3** 1-Dimensional Kernel Density Estimation plot of trip distances.



**Figure 4.4** 1-Dimensional Kernel Density Estimation plot of trip durations.

An ordinary least squares (OLS) regression is carried out to evaluate the change in log duration (y) as determined by the change in log trip distance (x). We see that there is coefficient of +0.386 log meters (coefficient tells you the size of the effect of x on y).

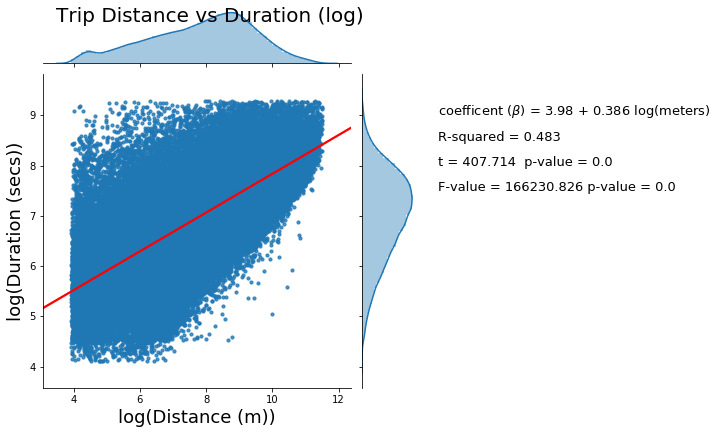
- relatively strong relationship, how well the line fits the data (R2=0.483)

- t-value is significant indicating difference in means,

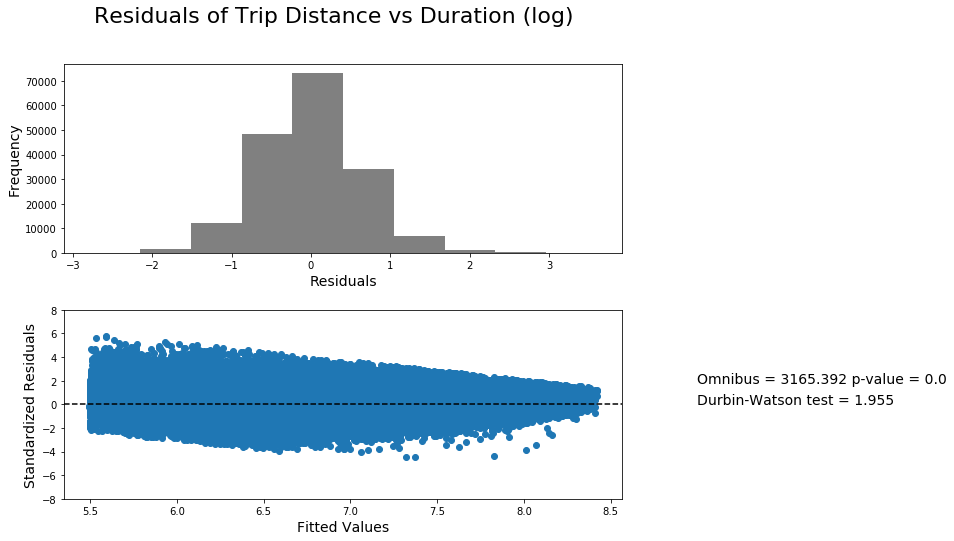
- F-value is probability the model is wrong (indicates the probability that all the coefficients in our regression output are actually zero)

- may need ref

Here, the residuals from the regression model were found to show non-normality (Omnibus below α=0.05) and



**Figure 4.X** OLS Regression results



**Figure 4.X** OLS Regression residuals

Breusch-Pagan test for Heteroscedasticity:

'B-P Test Statistic': 8757.14320669015, 'B-P Test p-value': 0.0, 'F-Statistic': 9210.327120693702, 'F-Test p-value': 0.0

Result shows: The data is heteroscedastic

“The Breusch-Pagan tests affirm that the residuals show no statistically significantly (p<0.05) heteroscedasticity, and thus are spatially random”

# Appendices

## Appendix 1 Notification not to apply for Ethical Approval

## Appendix 2 Python Script used for retrieval of Climatological data:

Code is available online at https://github.com/Thomasjkeel/MSc\_Dissertation under ‘. (This program consists of 500 lines written in Python computer language).  
*Description:* contains the code to create the figures and carry out the statistical analysis of this IGS.

# Appendix 3 Mean Direction and Distance Calculations

is calculated for each pair of points in each trip, the mean direction is calculated from the collection of these points using:

Calculations for mean direction in degrees:

*V\_east = magnitudes \* np.mean(np.sin(from\_dir \* np.pi/180))*

*V\_north = magnitudes \* np.mean(np.cos(from\_dir \* np.pi/180))*

*then..*

*mean\_dir = np.arctan2(V\_east, V\_north) \* 180/np.pi*

*mean\_dir = (360 + mean\_dir) % 360*

*mean\_dir = np.mean(mean\_dir)*

*Calculations for mean distance (magnitude) in degrees:*

*C = (1. / len(from\_dir)) \* (np.sum(np.cos(from\_dir \* np.pi/180)))*

*S = (1. / len(from\_dir)) \* (np.sum(np.sin(from\_dir \* np.pi/180)))*

*then…*

*resultant\_magnitude = (C\*\*2 + S\*\*2)\*(1./2.)*