Analysing travel times in London using Uber Data

Shane Flynn
National College of Ireland
Dublin, Ireland
x18122957@student.ncirl.ie

Martin Mohan
National College of Ireland
Dublin, Ireland
x18191339@student.ncirl.ie

Declan Moore
National College of Ireland
Dublin, Ireland
x18150713@student.ncirl.ie

Bolu Obitayo
National College of Ireland
Dublin, Ireland
x16138821@student.ncirl.ie

***Abstract*—What are the effects of bank holidays, weekends and weather on travelling times in a city like London? Using data from uber travelling times data were visualised and then analyzed to provide an understanding of the different predictors.**

***Index Terms*—uber, London, traffic**

I. Introduction

Uber Movement offers researchers a rich insight into travel times in major cities in a way that was previously unavailable [1] . This data offers the opportunity to city planners to reduce congestion, emissions, and improve road safety. Uber case studies in local areas have been successful in using this data, in Cincinnati a post pilot analysis has shown traffic crashes to be down 39% [2]. Our objective is to visualise this data and make it more accessible for analysis. We investigated how travel time in Camden London is affected by holidays and weather. Information on accidents, weather, and seasonality is displayed in a qlikview dashboard. Travel times were visualised using flourish and modelled using a linear regression and a random forest. We approached the visualisation as follows...

- IntroductionI
- Literature ReviewII
- Data Understanding III
- Data Preparation IV
- Visualisation using flourish V
- Visualisation using qlikview VI
- Modeling using regression and random forest VII
- Evaluation VIII
- Deployment X
- Conclusion and Future Work XI

II. Literature Review

London has over 10,000 taxis that are equipped with GPS tracking. These taxis generate huge volumes of journey trajectories daily. The journeys can provide urban planners with rich data to aid in improving city design. Taxicabs that allow GPS tracking of journeys allows urban planners to discover and investigate flawed urban planning and disjointed regions [3]. Disjointed regions are associated with subgraph patterns of region pairs with traffic problems, and regions with linking structure [4].

Previous studies of traffic behaviour relied on static predictions on traffic flow rather than a more realistic model of the distribution of human activity. Previous attempts to model travel behaviour relied on static models. Even fluid models that explored city structure, such as that created by Roth et al using data from the London Underground was not fully indicative of urban travel as underground travel routes are more constrained than road traffic.

Topological studies of cities, and studies using betweenness centrality measure of the nodes of street primal and dual networks have been found to be inferior models [5]. Using the spatial heterogeneity of human activities and the distance-decay law, has been found to be the best way to explain the observed traffic-flow distribution [6]. These activities, rather than the physical structure of the city, reveal the connection patterns within the city.

Traffic at different points in time allows a deeper understanding of city structure. By examining temporal graphs at different points in time to spatial graph of the city structure, traffic bottlenecks and rush hour traffic can identify how communities move and change through the day [7].

Real traffic flow changes throughout the day, and throughout the week. Traffic trips on workdays are more indicative of general traffic movement patterns as these journeys are more regular compared to journeys taken for entertainment on Fridays, weekends, and holidays.

Short journeys, which make up the majority of taxi journeys, are superior again in discerning travel patterns within a city. The short trips are between communities and hot spots within urban areas, whereas long trips ten to originate or end in airports and railway stations.

By using a temporal networks model, Liu et al. were able to measure changes in the centrality measures and community detection for modelling traffic flow through the static spatial network. This allowed them to pinpoint traffic bottlenecks at certain times of the day. This analysis city planners a superior model for future design, and for using simulated data for future changes [5].

By visualising human activity, using temporal and spatial data we aim to display human patterns of activity using GPS travel data from Uber, in a rich and easy to understand format.

III. Data Understanding

Uber have published Uber movement which details data of movements between different zones across cities served by Uber [8]. Using data from here we investigated how travel times in London varied depending on different independent variables. We obtained data such as weather from ???, holidays from ???

IV. Data Preparation

We combined Uber data with data from ??? and merged together using excel
Table I is an extract from the final csv file. The data for weekends,bank holidays, Christmas and Easter was modified using dummy encoding.

Date	TimeClass	Ave	TempKelvin	Rainfall	...
01/01/2016	RushMorning	3	276	0.71	...
02/01/2016	RushMorning	3	281	0.11	...
03/01/2016	RushMorning	3	278	0.08	...
04/01/2016	RushMorning	3	279	0.21	...
05/01/2016	RushMorning	3	280	0.10	...

TABLE I
SAMPLE OF UBER DATA

V. Visualisation using Flourish

Tuesdays were found to be the busiest days of the week for traffic, so we visualised how traffic varied over the year depending on different times of day using flourish.

A. Tower Hamlets to Camden

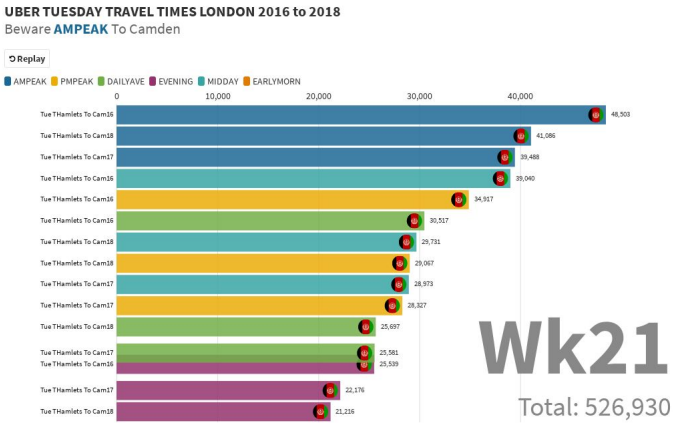


Fig. 1. Tower Hamlets to Camden https://public.flourish.studio/visualisation/945776/

B. Chelsea To Camden

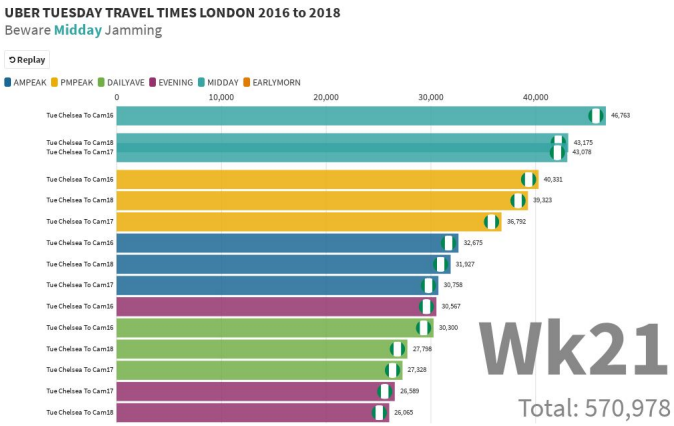
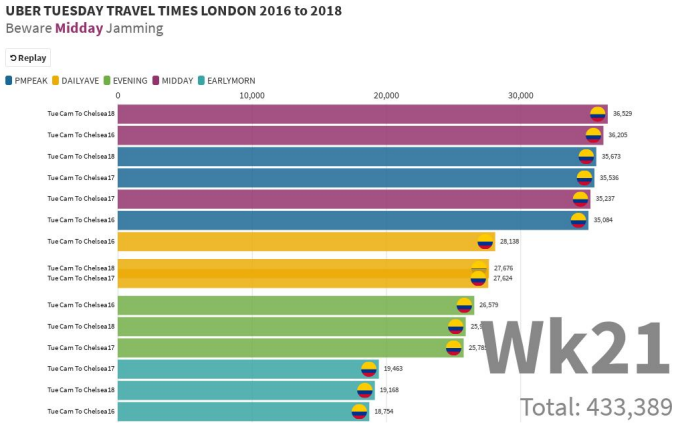


Fig. 2. Chelsea To Camden https://public.flourish.studio/visualisation/945870/

C. Camden To Chelsea



W

Fig. 3. Camden To Chelsea https://public.flourish.studio/visualisation/945883/

VI. Visualisation using QlikView Dashboard

QlikView was used to visualise the data see fig 4 There are two tabs on the dashboards: accident and traffic. There are filters on each tabs of the dashboard. Filters on the Vertical section of the dashboard include Season, Rush Hour, Rain Magnitude Snow Depth etc. The calendar filter is present on top of each page. The filter carries through each page of the dashboard.

Fig. 4. Dashboard homepage

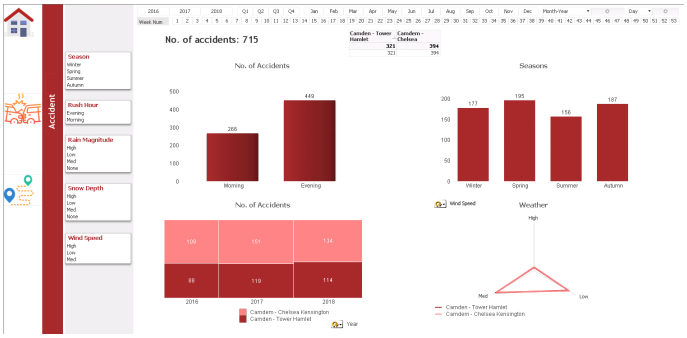


Fig. 5. Approximately 48% of the 715 accidents occurred between Camden and Tower Hamlets and 52% of accidents occurred between Camden and Chelsea.

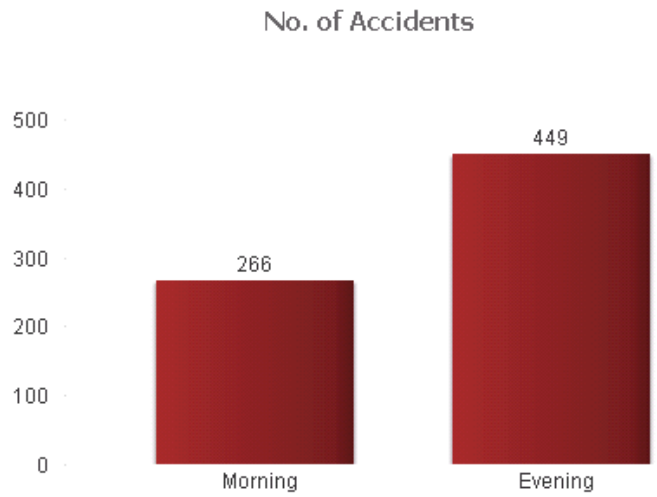


Fig. 6. Number of accidents: Morning vs Evening

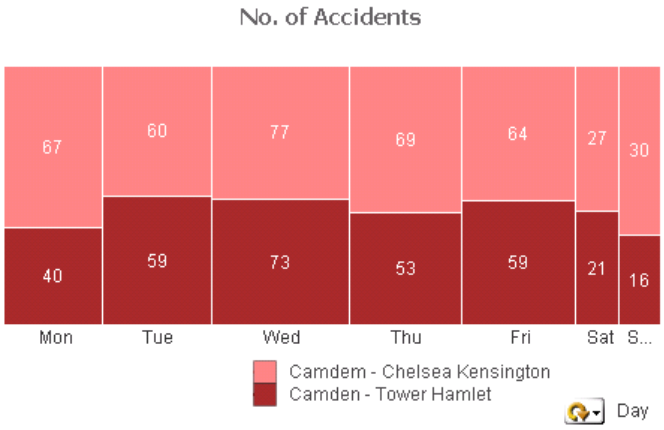


Fig. 7. Number of accidents: Day of week

A. *Accidents*

The accidents data captured were for Camden to Tower hamlet and Camden to Chelsea. There were 715 accidents in total between 2016 to 2018.

Majority (63%) of the accidents occurred in the evenings with majority of the accidents occurring between Wednesdays to Fridays.

B. *Commute Time*

There are two tabs within this page. One for the daily average and the other for the peak times. The average of the total time takes (seconds) was recorded.

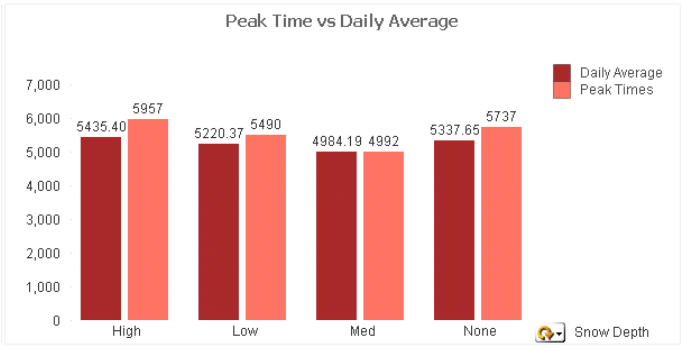


Fig. 8. Commute Time: Peak vs Daily Average

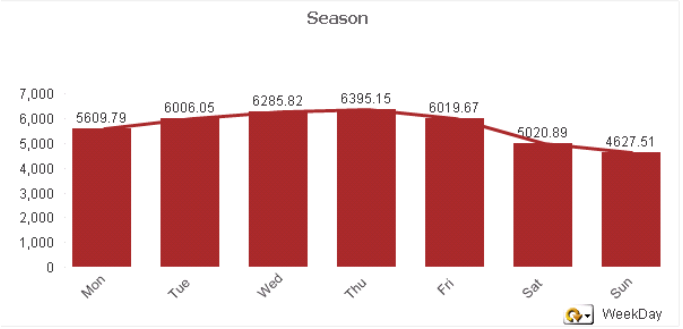


Fig. 9. Commute Time: Day of Week

1) Weather: The weather has a slight impact on commute times. e.g. During peak times, the average time taken to commute is 5 minutes quicker when the rain magnitude is high compared to when it is low.

2) Bank Holidays: On average, the commute time is faster on bank holidays for both the daily average and peak times. On bank holidays there are less cares out because most businesses are closed so there is no urgency to leave the house for most people.

3) College Holidays: On average, the commute time is faster when colleges are closed/on holidays for both the daily average and peak times.

4) Season: The longest and shortest commute time occurs during autumn and winter respectively.

5) Week Day: There is less traffic on weekends compared to weekdays, traffic on Wednesdays and Thursdays are the worst.

VII. Modelling using Machine Learning

Journey times were modelled based on the trips shown in table II. We found Camden to ChelseaKensington for morning rush peak was missing too many values to be useful so it was not included in the modelling.

ID	Origin	Destination	Peak
Peak_CTSecs_AM	Camden	Tower Hamlets	AM
**Peak_CK_AM*	*Camden*	*ChelseaKensington*	*AM**
Peak_TCSecs_AM	Tower Hamlets	Camden	AM
Peak_KCSecs_AM	ChelseaKensington	Camden	AM
Peak_CTSecs_PM	Camden	Tower Hamlets	PM
Peak_CK_PM_PM	Camden	ChelseaKensington	PM
Peak_TCSecs_PM	Tower Hamlets	Camden	PM
Peak_KCSecs_PM	ChelseaKensington	Camden	PM

TABLE II
Uber trips between London boroughs

The data obtained in table I was a mixture of continuous values (e.g. rainfall) and categorical values (e.g. Mon, Tues....). In order to predict travel times we chose predictors that were statistically significant by running a linear regression. Linear regression only works on continuous values or encoded values, so it was necessary to use dummy encoding on the categorical values see printout **??**.
Two continuous values were found to be statistically significant Rainfall and TempKelvin. School Holidays and ordinary working days Tuesday to Friday (OTues-Thurs,OFriday) were found to be statistically significant dummy encoded values.
Surprisingly snowfall and Mondays were not found to be statistically significant so they were not used as predictors when modelling the data.

```
Call:
lm(formula = Peak_CK ~ TempKelvin + Rainfall + SchoolHols +
    BHFriday + BHMonday + BHWeekend + OFriday + OMonday +
    OTues_Thurs + OWeekend + XEFriday + XEMonday +
    XETues_Thurs + XEWeekend, data = dtrain)

Residuals:
    Min      1Q  Median      3Q     Max
-469.77 -152.63  -25.79  111.49 2164.39

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    538.522    353.833   1.522  0.12841
TempKelvin       3.578      1.250   2.862  0.00433 **
Rainfall       109.313     38.348   2.851  0.00448 **
SchoolHolsYes -160.914     20.193  -7.969 5.45e-15 ***
BHFriday        83.503    110.380   0.757  0.44956
BHMonday      -171.928    101.342  -1.697  0.09017 .
BHWeekend       76.725     89.590   0.856  0.39203
OFriday        211.152     67.783   3.115  0.00190 **
OMonday         79.657     67.983   1.172  0.24165
OTues_Thurs    269.227     64.953   4.145 3.76e-05 ***
OWeekend       115.940     65.346   1.774  0.07640 .
XEFriday        74.739    108.759   0.687  0.49216
XEMonday         3.309    104.048   0.032  0.97463
XETues_Thurs    95.209    115.034   0.828  0.40811
XEWeekend           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 237.4 on 808 degrees of freedom
Multiple R-squared:  0.2326,Adjusted R-squared:  0.2202
F-statistic: 18.83 on 13 and 808 DF,  p-value: < 2.2e-16
```

Only statistically significant predictors were used in modelling. The data was split into 75% training data and 25% test data. We used two models **linear regression** and **random forest** to predict travel times using the training data. The test data was used to predict values and the results are shown in section VIII

VIII. Evaluation

A. Evaluation of Linear regression/Random Forest

We wanted to predict travel times using two different models **linear regression** and **random forest**. MAE **Mean Absolute Error** is used to compare predicted values against actual values (when both values are continuous). The accuracy of the predicted travel times using both methods is shown in table III. The table shows that random forest was a better predictor (smaller MAE) of travel times than linear regression.
E.g. in the morning rush hour Camden to Tower Hamlets takes on average 1339 seconds. Using linear regression we can predict travel time to $+/-$ 290 seconds but using random forest that error in prediction falls to $+/-$ 193 seconds.

ID	Average(s)	MAE (LM)	MAE (RF)	Peak
Peak_CTSecs_AM	1339	290	193	AM
Peak_TCSecs_AM	1673	505	331	AM
Peak_KCSecs_AM	1367	345	223	AM
Peak_CTSecs_PM	1852	313	202	PM
Peak_CK_PM_PM	1708	227	181	PM
Peak_TCSecs_PM	1466	210	160	PM
Peak_KCSecs_PM	1653	222	174	PM

TABLE III
Uber trips: **MAE** (linear regression lm or random forest rf)

???

IX. Infographic

???...

X. Deployment

Commuter behaviour may be used to aid consumers as demonstrated with the new DART service in Dublin used to help commuters avoid crowded trains [9] ... ???

XI. Conclusion and Future Work

???...

References

[1] Uber. *Uber Movement: Let's find smarter ways forward, together.* URL: https://movement.uber.com/?lang=es-CO (visited on 11/16/2019).

[2] Uber Movement Team. *Improving Road Safety in Cincinnati's Northside neighborhood*. Medium. May 14, 2019. URL: https : / / medium . com / uber - movement / oki - 8b77c95bb368 (visited on 11/16/2019).

[3] Yu Zheng et al. “Urban computing with taxicabs”. In: *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp ’11*. the 13th international conference. event-place: Beijing, China. ACM Press, 2011, p. 89. ISBN: 978-1-4503-0630-0. DOI: 10.1145/2030112.2030126. URL: http://dl.acm.org/citation.cfm?doid=2030112.2030126 (visited on 10/30/2019).

[4] Alasdair Turner. “From Axial to Road-Centre Lines: A New Representation for Space Syntax and a New Model of Route Choice for Transport Network Analysis”. In: *Environment and Planning B: Planning and Design* 34 (May 1, 2007), pp. 539–555. DOI: 10.1068/b32067.

[5] Song Gao et al. “Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality”. In: *Environment and Planning B: Planning and Design* 40.1 (Feb. 1, 2013), pp. 135–153. ISSN: 0265-8135. DOI: 10.1068/b38141. URL: https://doi.org/10.1068/b38141 (visited on 10/31/2019).

[6] Camille Roth et al. “Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows”. In: *PLOS ONE* 6.1 (Jan. 7, 2011), e15923. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0015923. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015923 (visited on 10/31/2019).

[7] Mackenzie Pearson, Javier Sagastuy, and Sofia Samaniego. “Traffic Flow Analysis Using Uber Movement Data”. In: (), p. 11.

[8] Uber. *Uber Movement: Let’s find smarter ways forward, together.* URL: https://movement.uber.com/?lang=en-US (visited on 11/04/2019).

[9] LovinDublin.com. *New Service Tells DART Commuters Exactly How Crowded Their Trains Are Every Morning*. LovinDublin.com. URL: https://lovindublin.com/dublin/new-service-tells-dart-commuters-exactly-how-crowded-their-trains-are-every-morning (visited on 11/24/2019).