INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

SURGE – 2022

Project Report

‘Exploratory Analysis of Travel Times in Delhi using UBER Movement data’

Submitted by

Chinmay Joshi

Application no.: 2230207

Department of Civil Engineering, IITK

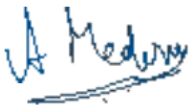
Under the guidance of

Prof. Aditya Medury

Department of Civil engineering, IITK

CERTIFICATE

This is to certify that the project titled **'Exploratory Analysis of travel times in Delhi using UBER Movement Data'** submitted by **Chinmay Joshi** (2230207) as a part of Summer Undergraduate Research and Graduate Excellence (**SURGE**) 2022 offered by the Indian Institute of Technology, Kanpur, is a Bonafede record of the work done by him under my guidance and supervision at the Indian Institute of Technology, Kanpur from 23rd May 2022 to 29th July 2022.



Prof. Aditya Medury

Department of Civil Engineering

IIT Kanpur

ACKNOWLEGDEMENTS

Firstly, I express my gratitude towards IIT Kanpur for providing such a brilliant opportunity to work under the direct guidance of professors from our esteemed institute. Due to this opportunity, I was introduced to new ways of thinking and managing information.

Secondly, I was constantly guided and kept on track by my mentor Prof. Aditya Medury, who also introduced to me the many ways to develop a simple question into a scientific inquiry and nudged me along as I was introduced to cutting edge tools and functionalities.

Finally, my parents also helped me by providing a distraction free environment wherein I could work efficiently.

Hence, I hereby thank everyone whose help and guidance were integral to the making of this project.

ABSTRACT

The **yearly trends** in the **travel-times** of the **Movement data** collected by UBER from 2016 quarter 1 to 2019 quarter 3 are **statistically tested**. In this project, specific **origin-destination pair travel-times data** is **aggregated** to compare their **yearly distributions** using R. Then **statistical tests** to validate the changes are carried out and **confidence intervals** for the changes are established.

KEYWORDS*: Travel-times, Movement data, statistical analysis.*

CONTENTS

INTRODUCTION

The need to analyse movement data is evident from the increasing utility of making more efficient transportation infrastructure as urbanisation and standard of living leads to higher vehicle counts, congestion and fuel consumption across cities all over the world. When quantifying the efficiency of a transportation network, the time it takes to travel from location A to location B is the simplest metric to collect and utilise.

In most scenarios, the travel time is not a direct function of the Pythagorean distance between the origin and destination locations, but rather the distance along the network that connects those two points. Hence, travel times are generally collected between zones or small geospatial polygons and not exact locations.

For each zone, a unique id gives the location of the zone and hence an id pair uniquely identifies a 'trip' from the origin zone (first id) to the destination zone. In this manner the travel times for a collection of zones is stored as an origin-destination or OD matrix.

Various aggregators across the world report about various conditions of the transportation infrastructure, like traffic, congestion on roads, emissions etc. This is achieved by using statistical modelling and testing on various movement data to find and interpret patterns across time and space. Within this study, the focus is specifically on travel-times data of New-Delhi as collected by UBER from 2016 to 2019. This data shall be used to find how the travel-times within the city changed across years, whether the distribution of the travel times changed, or stayed uniform, and was there a general trend that could be statistically tested.

BACKGROUND

UBER movement data [1] is where UBER publishes the travel-times data it collects from its trips and aggregates them into daily, weekly and monthly formats. This data is not collected for location to location, but for zone to zone. Here a zone is a geospatial demarcation uniquely identified by an index. For New Delhi [2], the number of zones is 290, which means there can be 84100 unique origin-zone destination-zone pair (OD pair) trip categories in the data. But due to the methodology of data collection used, a minority of the O-D pairs do not get recorded due to insufficient number of trips for anonymization. [1.1]

Another factor of uncertainty comes from the fact that the trips categorized under the same OD pair may not have been conducted over the same distances as the trip might have started from any location within the geospatial boundary of origin zone and ended similarly at any location within the destination zone. The statistical analysis within this project assumes that due to the requirements of anonymization, any travel-time within the UBER movement data is an aggregate of a sufficiently high number of trips and that it approaches a central tendency of the travel-time of the underlying network connecting the two zones.

Finally, the dataset did not contain the last quarter of the year 2019 most logically due to not having enough trips during the onset of the COVID pandemic. Hence for this analysis, the last quarter of years 2016, 2017 and 2018 have also been omitted for consistent comparison of all the years.

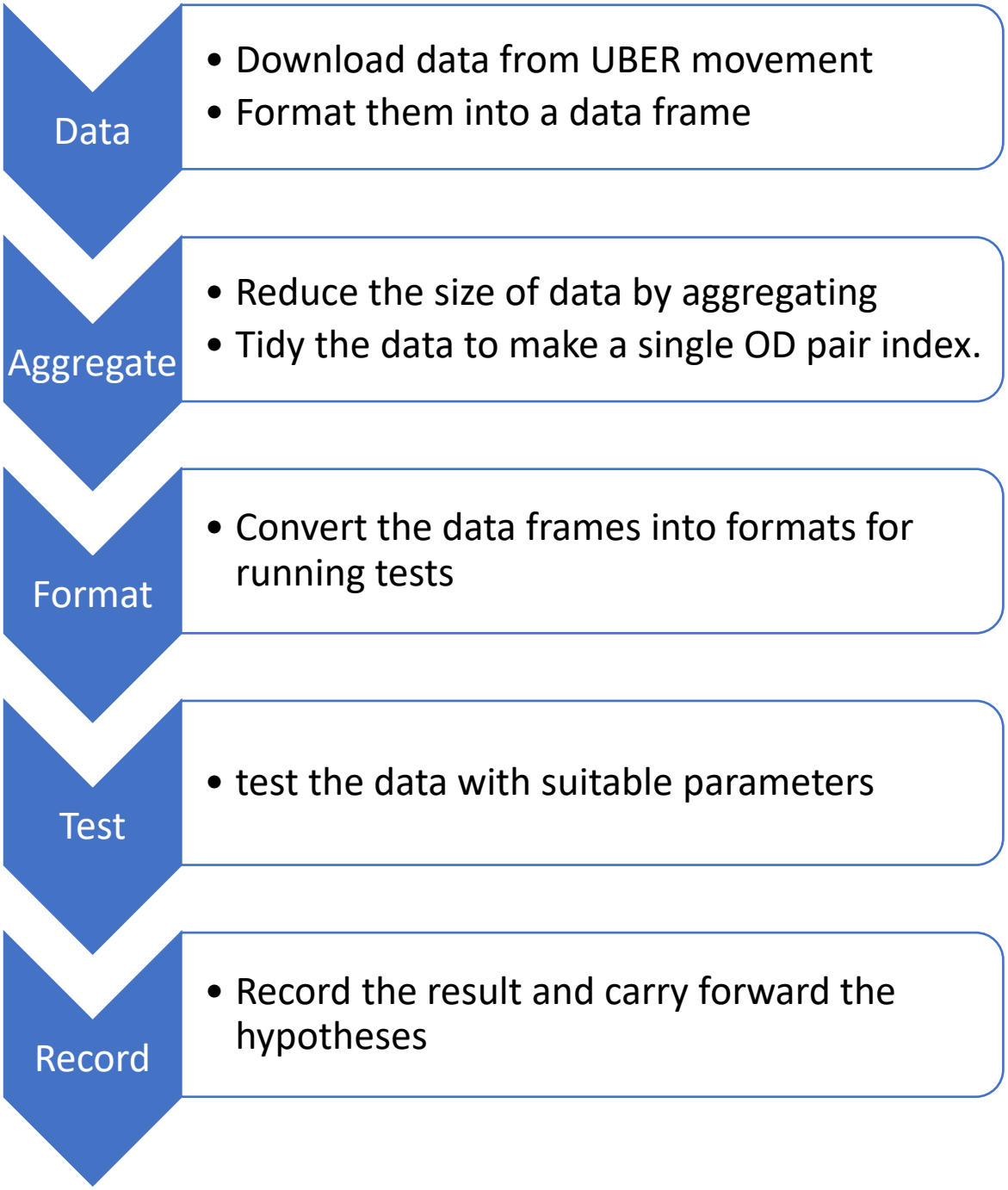
With these prerequisites for the final datasets to carry out analysis on, the R programming language was used to format, aggregate and filter the data to make four yearly data-frames with 66945 unique OD pairs.

Now considering the fact that 17155 or more than 20% of all possible OD pairs have been excluded from the analysis, the travel-time statistic may not be representative of the true value on road. But this can be excused over the fact that only the OD pairs that could be consistently anonymized over the four years have been considered, i.e., the number of trips conducted over these OD pairs are representative of the majority of the trips mediated by UBER and thus it can be assumed for this study that the excluded OD pairs do not have a significant effect on the aggregated statistics of the overall yearly samples.

On the subject of pre-recorded trends in the travel-times data, the congestion levels of New Delhi have been recorded to show a sharp decline from 2019 (56%) to 2020 (47%) and then showed a gradual increase in 2021 (48%) according to the aggregator TomTom.[3] The congestion level of a city is calculated by the aggregator as the percentage difference between travel-times in current conditions in proportion to the travel-time in a baseline non-congested scenario.[3]

Another report also states that congestion levels decreased from 2018 to 2019 by 2% [4] and 2017 to 2018 by 4% [5]. This begs the question; what trend was shown by the distribution of the travel times of the city across the years. The hypothesis for this study is that **the travel-times of the New Delhi city decreased over the years.**

METHODOLOGY



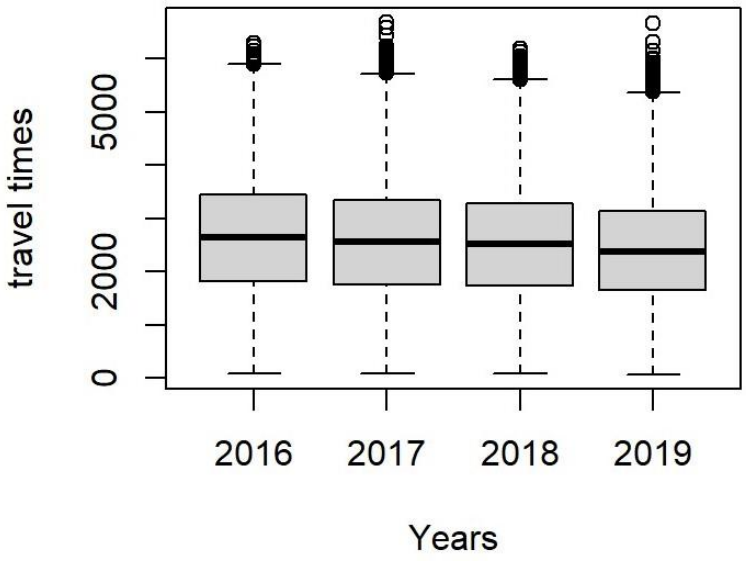
This is the basic methodology that has been used in this project. Before starting with running tests, we shall also visualize the data separately and try to find visual plausibility for the hypotheses.

DATA

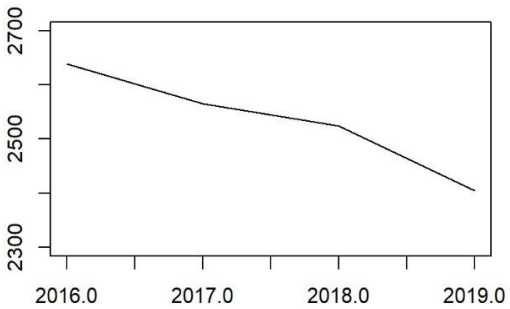
VISUALISATIONS

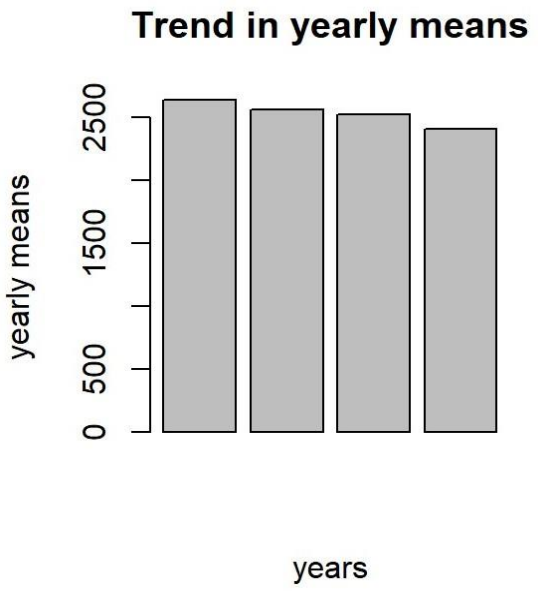
In order to better understand the data, several visualisations are provided by the R programming language. As a first step towards finding trends within the years, let us compare the four samples side-by-side.

The boxplot of the four yearly dataset gives a simple comparison between the distribution of the samples.



It can be seen that the distributions do not differ much in absolute terms, even though there is a slight shift in the positions over the years. Perhaps, plotting a central tendency such as the mean could give greater insight into the magnitude and direction of the shifts in the distribution.





Here, even in absolute terms the decline is apparent, but to show if the decline is of significance, we need to show that the underlying distributions of the datasets are different.

STATISTICAL TESTS

To show that the underlying distributions of the yearly samples are different, the Kolmogorov-Smirnov test [6.1, 6.2] can give a metric of the probability that the two samples come from the same distribution. If the probability/p-value of the test comes below significance (<0.05) we can conclude that the samples do not come from the same distribution.

Another important piece of information about the sample distribution is if it can be considered to come from a normal distribution. To find the plausibility of this, we shall use the Jarque-Bera Test for normality. [7] When we can assume the distribution to be normal, we can use the paired sample t-test [8] for finding the directionality of the change or the paired sample Wilcoxon test [9] if normality cannot be presumed.

RESULTS

H1: The distributions stayed uniform over the yearly sample pairs: 2016-2017, 2017-2018, 2018-2019.

Test used: Two sample Kolmogorov-Smirnov test [6.1, 6.2]

Alternate Hypothesis: The samples come from different underlying distributions

Results:

YEARS	P-VALUE	INFERENCE
16 -17	< 2.2e-16	Alternate: Distributions differ
17 -18	2.824e-09	Alternate: Distributions differ
18 - 19	< 2.2e-16	Alternate: Distributions differ

H2: The distribution of the travel times in the yearly samples is normal.

Test used: Jarque-Bera test for normality [7]

Alternate Hypothesis: the sample distribution deviates significantly from normality.

Results:

YEAR	P-VALUE	INFERENCE
2016	<2.2e-16	Alternate: significant deviation
2017	<2.2e-16	Alternate: significant deviation
2018	<2.2e-16	Alternate: significant deviation
2019	<2.2e-16	Alternate: significant deviation

*This means that the sample distributions deviate significantly from a normal distribution. Hence, normality cannot be inferred from the test.

**For non-normal sample distributions, the Wilcoxon signed rank test can give us a measure for the significance of the direction of the variation among the yearly sample pairs.

H3: The yearly sample of 2016 contains smaller travel times than that of 2017; and so on for (2017, 2018) and (2018, 2019).

Test used: Paired sample Wilcoxon signed rank test [9] **

Alternate Hypothesis: The yearly sample of 2016 contains greater travel times than that of 2017; and so on for (2017, 2018) and (2018, 2019).

YEARS	P-VALUE	INFERENCE
16 -17	< 2.2e-16	Alternate
17 -18	< 2.2e-16	Alternate
18 - 19	< 2.2e-16	Alternate

Hence, the decline in travel times across the OD pairs is statistically significant.

But it would be much more conclusive if a confidence interval could be established for the drops in travel times across years. Unfortunately, that shall require us to assume a distribution for the travel times. Hence, let us assume that the samples follow a loosely normal distribution. This makes computation much simpler.

Even though we have shown that the samples do not follow the normal distributions, we can take the assumption since, the distributions do follow a bell-like curve and the skew [10] for the distributions is well within the acceptable range.

YEAR	PEARSON'S COEFFICIENT OF SKEWNESS
2016	0.01235329 < 0.4
2017	0.009026067 < 0.4
2018	-0.0001721346 > -0.4
2019	0.06678994 < 0.4

Visually,

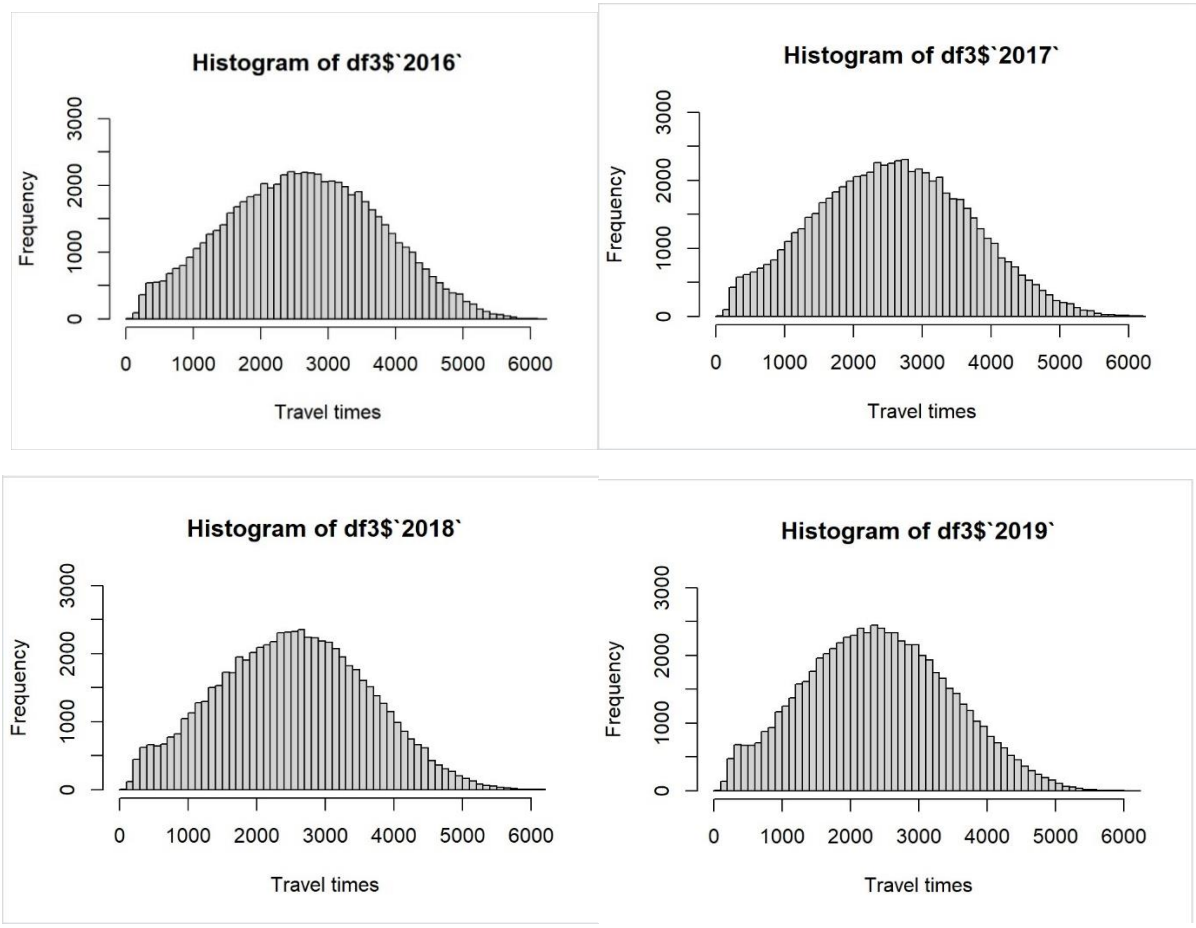


Fig: *the distributions of travel times show an approximate bell shape.*

Results from the t-test [8]:

YEAR PAIRS	95% CONFIDENCE INTERVALS	MEAN DIFFERENCE
16 – 17	[72.91330, 75.12967]	74.02149
17 – 18	[39.49212, 41.39311]	40.44262
18 – 19	[117.848, 120.875]	119.3615

Hence, we can conclude that the travel times in the Delhi city decreased significantly across the years from 2016 to 2019. Here, the exact numbers of decrease cannot be pinpointed and the confidence intervals are only approximate due to the normality assumption. But all the statistical tests point unequivocally towards decreasing travel times.

DISCUSSION AND CONCLUSION

The study concluded that the travel-times within the New-Delhi city decreased in general. This observation is counter-intuitive to say the least, considering that the traffic flow has been observed to increase across the world. There could be many factors that could lead to lower travel-times such as car speeds, improving transportation networks, better traffic-modelling and re-routing algorithms etc. These could be some factors that could be investigated for further study for the causality of this trend.

Another aspect of further investigation could be correlating the decreasing travel times to specific routes on the network to find which modifications to the transportation infrastructure led to better travel-times and which did not.

With the factors for further investigations discussed, the following study has been completed.

SUMMARY

The data from UBER movement was formatted and analysed in R-studio and subsequently visualised by various plots. The plots showed a downward trend in the travel-times of the New-Delhi city and the statistical significance of these findings was acquired through hypothesis testing methods. The test results supported the trends seen visually and thus confidence intervals were established for these trends. The study was concluded by discussing the possible causes of the trends and further investigations in that direction.

REFERENCES

[1] UBER Movement: https://movement.uber.com/?lang=hi-IN

[1.1] UBER Movement: Travel time calculation methodology: https://movement.uber.com/_static/c9bce307d99643c3.pdf

[2] UBER data, New Delhi: https://movement.uber.com/explore/new_delhi/travel-times/query?si=101&ti=&ag=wards&dt[tpb]=ALL_DAY&dt[wd;]=1,2,3,4,5,6,7&dt[dr][sd]=2020-03-01&dt[dr][ed]=2020-03-31&cd=&sa;=&sdn=&lang=en-US

[3] Tomtom (n. d.): https://www.tomtom.com/en_gb/traffic-index/new-delhi-traffic/ (referenced by [4], [5])

[4] Orissa diary, 2020. 'New Delhi is among the most traffic congested city in the world': https://orissadiary.com/new-delhi-is-amongst-the-most-traffic-congested-city-in-world-tomtom-traffic-index-2019/

[5] Mint, 2019 'Mumbai's traffic flow worst in the world, Delhi at fourth spot': https://www.livemint.com/news/india/mumbai-s-traffic-flow-worst-in-world-delhi-at-fourth-spot-says-report-1559673875430.html

[6.1] R documentation(n.d.); Kolmogorov-Smirnov test: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ks.test.html

[6.2] 'ks.test: Kolmogorov-Smirnov Tests' in RDocmentation: https://www.rdocumentation.org/packages/dgof/versions/1.4/topics/ks.test

[7] 'jarque.bera.test: Jarque-Bera Test for Normality' in RDocumentation: https://www.rdocumentation.org/packages/tsoutliers/versions/0.3/topics/jarque.bera.test

[8] 't.test: Student's t-Test' in RDocumentation: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

[9]'wilcox.test: Wilcoxon Rank Sum and Signed Rank Tests' in RDocumentation: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test

[10] Greg Deckler (2019), Microsoft Power BI Community, 'Pearson's Coefficient of skewness': https://community.powerbi.com/t5/Quick-Measures-Gallery/Pearson-s-Coefficient-of-Skewness/m-p/623533