Coursework submission for Transport Data Science (TRAN5340M)

Level of walking in Leeds & Pedestrian safety

Student 201484781

Introduction

In present age,transportation engineers and researchers requires an understanding of data and databases in order to establish quantitative, qualitative and empirical facts to validate and analyze different network, paths and modes of transportation.It's advantage over traditional methods is visualization of findings, which strengthens the credibility and legitimacy of the work.

Walking as we all know have various health benefits strengthens hearts, help in fat reduction eases joint stiffness, improve mood and boost energy. NHS recommends a daily brisk walk of 10 min or normal walk of 45 min for a healthy lifestyle. In the survey done in 2015,from the Department for Transport, Leeds was country's top city for increased walking. Also a survey was conducted by Living street in 2017, and 54% Leeds city residence felt safe walking and 72% found the quality of street good.61% said good pavements and 49% Ease of walks in parks.

1.1 Scope

The primary focus of the research is to find the level of walking in Leeds city and we would also analyze how do walking levels relate to pedestrian safety? A pedestrian, is any person on foot, walking, running, jogging, hiking.[https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811748]

We would do so by analyzing the different factors on which walking level is concentrated or less across the Leeds city and compare it with stats19 causality data to understand if the safety of pedestrians is one the reasons of less or more level of walking.

1.2 Area of study

Our area of study would be Leeds city.It is the largest city in the county of West Yorkshire, England, occupying 551.7 square kilometers of land. It is part of West Yorkshire county which has a population of 2.2 million (2011 Census), occupying an area of 2,030 km^2. It thus has a relatively high population density of 1,084 people per km^2 who are unevenly distributed between the 5 Local Authorities of Leeds, Bradford, Wakefield, Calderdale and Kirklees.Leeds and Bradford, have historically low levels of active travel and heavily car orientated urban plans.

Based on the 2011 census, 12.3% of the surveyed population walk to work in Leeds.[https://www.westyorks-ca.gov.uk/media/2847/transport-strategy-evidence-base.pdf] this compared to other modes of transportation, car/van, public transport is less.(18.7%) Car continues to be pre dominant mode of transportation. People generally prefer walking over other modes of transport if the distance is shorter and if it is relevant to commercial area.

If we discuss the safety of pedestrians over the years,the number of casualties according to stats19in Leeds city is- In the below figure 1,table we can see the casualties caused to pedestrians over the period of 13 years in Leeds city. The death rate was by far the lowest in 2013 and highest in 2012.

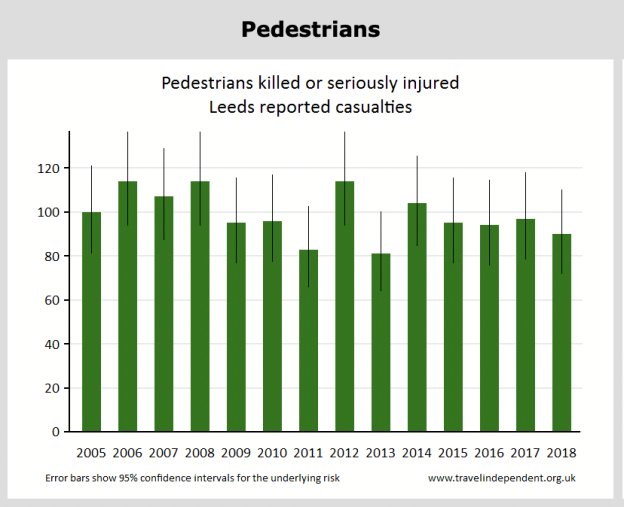

Figure 1: Leeds pedestrian casualtiy shown over period of years

1.3 Datasets

The main dataset for this research is 'stats19' package, this package contains all the Great Britain's official road traffic casualty database, we will use the casualties and crashes data from this package and join in with data obtained from `pct` Origin-destination data of Leeds developed on 2011 census data was obtained from the `pct` package to determine walking rate in Leeds and different routes wise % of walking. PCT provides us data for different routes and regions across a particular area for different modes of transportation. Stats19 data would help us analyse the casualities and pct data would help us find walking perecenatge in Leeds.

2 Data preparation

2.1 Installing datasets

We will download stats19 (2018) dataset and pct dataset for the region = 'West-yorkshire'

Once data is download, we will install relevant libraries. Like tidyverse, tmap, dplyr, ggplot,sf. All the relevant code is included in the zip.

2.2 Data preparation

Before analyzing the data, it was necessary to first understand the raw datasets with respect to our analysis and pre-process them into forms suitable for subsequent analyses. As we can see stats19 contains alot of information which are for all types of casualties, based on mode of transportation and casualty type. Since for our analysis we would only require 'Pedestrian' as casualty type, we would filter out data accordingly. Also for both PCT and stats19 we would filter the location by 'Leeds'

2.3 Data cleaning

For our data from stats19, we can see that we have 4 values in 'Time' column which are null and 921 values in 'second road class' column as NA. Since for our analysis we do not need second road class, we will proceed to remove the column and since the columns with time as NA is pretty small compared to the dataset, we will remove it as well. After cleaning stats19 dataset for 2018, we will create a data-frame crashes_leeds, which will have data from 'accidents' and 'casualties' from stats19/2018 data.We filtered it

only for 'Leeds' since that's our city of analysis. We then did left join for both crashes and casualties on the basis of 'accident_index'.The accident records for Leeds were filtered using the filter() function. Once we had all the crash data, since we are focusing on the safety and causalities of pedestrians alone, we filtered the crash and causality data for casualty_class=Pedestrians using fliter() function.

For our data from PCT dataset, for zone='west-yorkshire', we filtered it only for Leeds data for routes and then sanity checked for any missing values. Since none of the relevant columns had missing or NA values, it didn't need any cleanup.

For the purpose of understanding the graphs better, we will rename certain column values, like lighting conditions and weather conditions analyzed in section 3.2.4

3 Data Exploration

Now that we have our data with us, we will initialize our analysis- Here is a map of Leeds city[https://www.worldmap1.com/map/leeds-map] next to the walking plotted map of Leeds.Once we have the cleaned data, we need to combine the two datasets, PCT zone data with stats19 data to do further deep dive. We will do so by making first crashes data from stats19 to SF object and same for zone data from pct, then on the basis of a key we will join them in a single sf object.

Method- we are taking the geo_code column data from zones, matches it to the geometry column in crashes_sf and then joins it to the crashes that have occurred in those geo_codes. The matched, joined geo_code is a new column in the zone_joined dataset. We now know the administrative geo_code in which each crash occured.[https://itsleeds.github.io/rrsrr/space.html#geographic-joins]

We will first convert our crash data to a sf object (spatial class) We will subset the zones (pct data for Leeds area) that contain features in crashes (stats19 cleaned data for 2018, Leeds) using the notation [], and so we can visualize our geo location Leeds and different casualty. We are converting stats19 data and pct data in sf object because- mapping is a functionality we can utilize to analyze our crashes and compare it with pedestrian data and understand it.

3.1 Walking percentage across Leeds

We will filter out 'Leeds' from 'West Yorkshire' county and analyze the desired routes. We can see via our map, the percentage vise value of walking in the city Leeds, the blue shows the least percentage walking and yellow shows the maximum. As we can see from the map that majority of walking is below 20%. The maximum walking % is concentrated around one area, which is city center area, where major shopping and financial quarters are there. Which tells us majority of people prefer walking in the busy district. This is based on distance less than 3kms, so short walks across for work and shopping is preferred in the city.

We will use the method ggplot2(), filter(), sum() to plot and summarize our data. We have a general view of Leeds city in figure 2, [@http://www.travelindependent.org.uk/area_049.html] and we have plotted the routes and LSOA zones via our pct package for 'Leeds' city.As we can see already by the graph, walking parameters are more frequent in city center area. Both the geographical plotting and the routes across the city (via pct_desire_lines) corresponds to the finding.

We will use get_pct_zones() and get_pct_lines() function from the pct package to get our Leeds zones walking %.We will use the default 'lsoa' as geography variable.Also we will put purpose as default commute.The percentage of people walking to that of total mode of traveling is 34%. Whereas if we compare this percentage to bicycle has 27%. If we summarize our lines across the city we can see that mean average across the city for small distance is 2 kms and for distance above 3km it's is 5.2 kms.

In figure 3,We are comparing the number of trips taken by the residents across the city. We will plot is based on distance covered (0,1,2,3 and 4).We would visualize this based on distance covered, mode of transportation- car, bicycle, foot or any other. We used ggplot() method to plot the different modes with respect to distance covered in trips.We can also visualize the walking percentage by leaflet map view figure

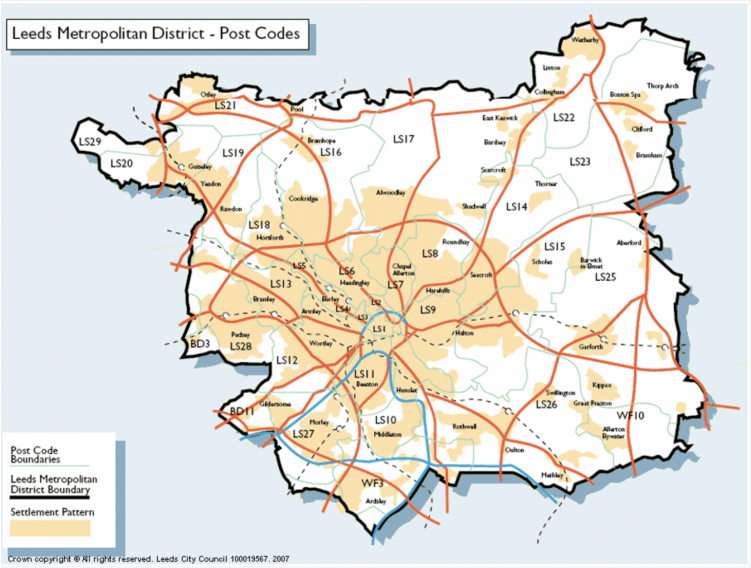

Figure 2: Leeds city map

4, which is an interactive view of the map, with Leeds as our zone and we are plotting desired lines (<3km, for calculating walking %) as we can see it's highly concentrated in the city center area with majority % of walkers are there. The full interactive map in available in RMD file.

We are using the interactive map under leaflet library to show the desire lines for Leeds and foot percentage.As you can see the amount of people walking less distance is much more than distance over 3kms. We calculated the weights for the map by taking sum of all walking lines over mean of overall routes five times.

In figure 6,We earlier calculated that if we only consider small distance (<=3km) then 34% people in Leeds prefer walking. Now, If we see the graph above and narrow our analysis, we can say that for Leeds city area, if the distance is between 0-1 kms maximum people prefer walking and as the distance increase the percentage decreases. Bicycle has a constant stand across the kms, as be it 0-1 or 2-3 kms almost similar number of people prefer bicycle trips. Whereas car trips are more obvious choice when distance increases.

Discussion

As per the analysis of the walking data, we can see that people prefer walking around city center area (Shopping complex and major eateries) majorly though the walking percentage here could mean people prefer walking more for leisure, shopping than work.People do prefer walking across parks as well. Shown in graph for example for a park, North east area of Leeds (shown in purple) but it's less. So walking for health in park is also not on top reasons to walk. And if we compare walking percentage in terms of distance, the lesser the distance the more people prefer walking. As distance increases, walking percentage goes down and other modes prevail.

Distance plays an important factor while mode of transportation is decided and the activity with respect to area. Now we will try and understand if walking percentage is driven by casualties across the city for pedestrians.

3.2 Pedestrian safety

As explained before, we have crashes data filtered at "Leeds" city level from stats19, 2018 data set. We will first try to understand how many foot-walkers meet casualties. We will use sf library (spatial) to convert stats19 cleaned crash data and pct zones data into sf objects to add geographical nuance to it.

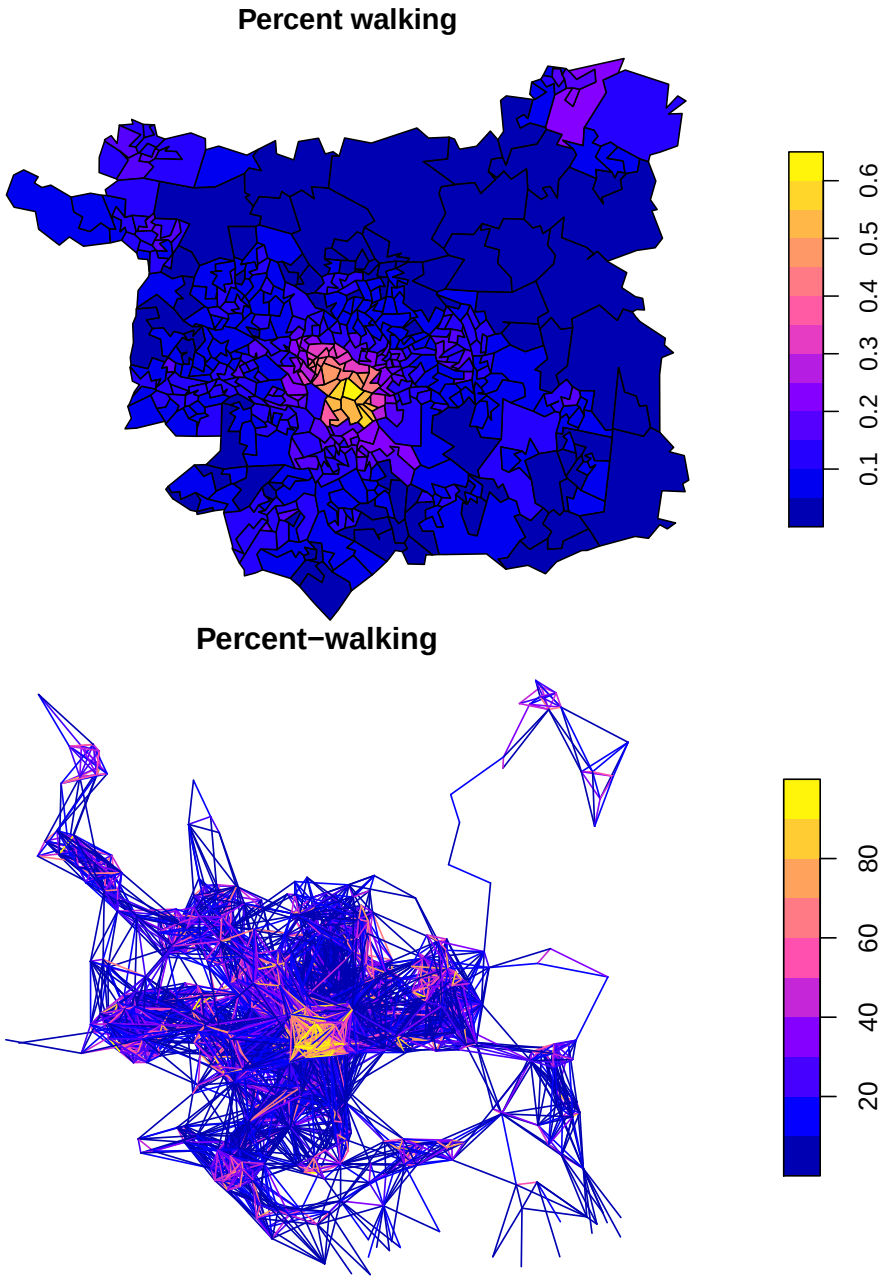

Figure 3: The desire routes and desires lines across Leeds city for walking percenatge over foot for disatnce <3 kms

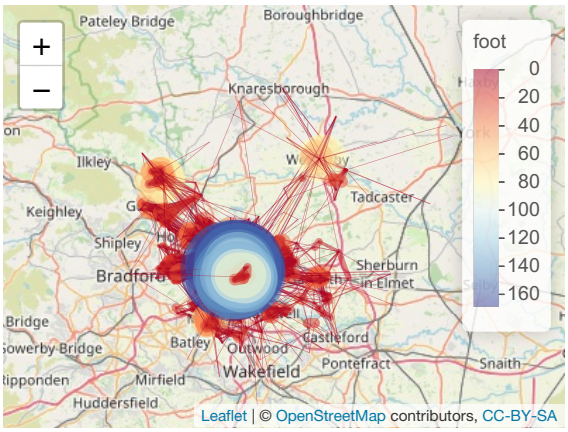

Figure 4: Leaflet map view of the on foot journey across Leeds and where it's more concentrated

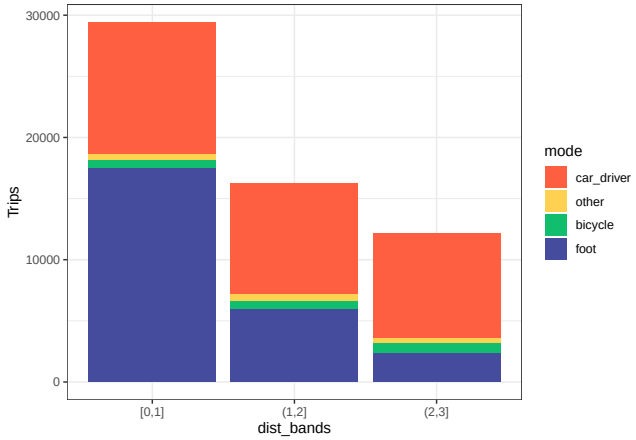

Figure 5: Comparing different modes of travel across the city

Once we create the sf object, we would use tmap() function to plot the Leeds city geographically, with dots showing the crashes happening for pedestrian.This we have shown in figure 7.

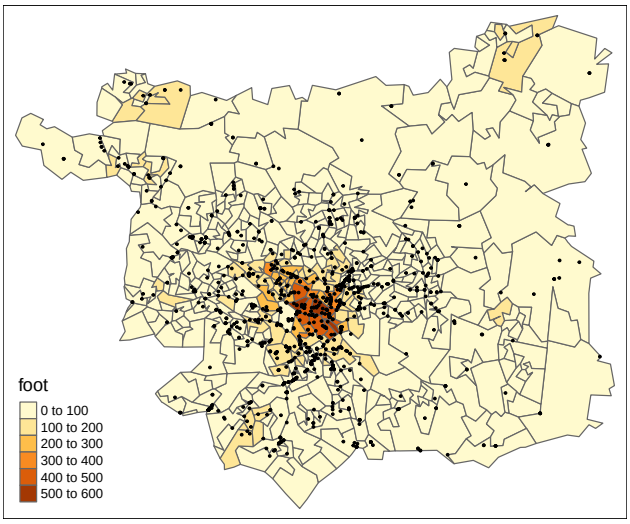

Figure 6: Geographically plotting pedestrian casualty over Leeds city

If we see this data, this corresponds to the walking level we saw in previous section and the crashes are frequent in areas of more walking for obvious reasons.(More people on foot in this region). The more we move out in the city the less we see casualties for pedestrians. So we need to understand what factors derive this casualty percentage

3.2.1 Sex of casualtiy vs Age of casuality

We will use geom_histogram function of ggplot() to plot a comparison between age of casualties and gender. If we see in the graph below in figure 8, the data for both male and female is skewed (right-sided) with peak for male at 14-15 age and for female less than 15 years. This suggests that young children are a target for such accidents. There are more males falling as victim for accidents than women, this doesn't mean women are safe walker or are not getting into accidents, this could also means less females walk. If we see for data for age 70+ the male and female are almost equally susceptible to fatalities. A total of 58% casualties are of male and 41% are for female. Children under 15 years are 7.8% of total casualties and elder above 50 are also 7.4%. Rest majority casualties are spread across the ages of 16-49 years.

If we see the ggboxplot() graph with casualties and age band, the most fatal accidents happens for adults of age 30-50 and slight severity for young adults of age 25. The fatal box is the tallest, signifying more data in fatal category.Serious and slight casualty are of almost same range with almost the same median age. All the boxes don't have any outliers showing that most of the pedstrian casualties are centric to median.

3.2.2 Time of Casualties

in figure 9,We are using ggplot and geompoint to plot the time at which most frequent accidents happened for pedestrians over the period of 2018.We will see this over 4 time periods, early morning, afternoon, evening, night. If we see that during the break time (Easter break) the number of casualties are in general less than other months, suggesting less traffic. September shows more accidents happening, which also coincides with University opening.

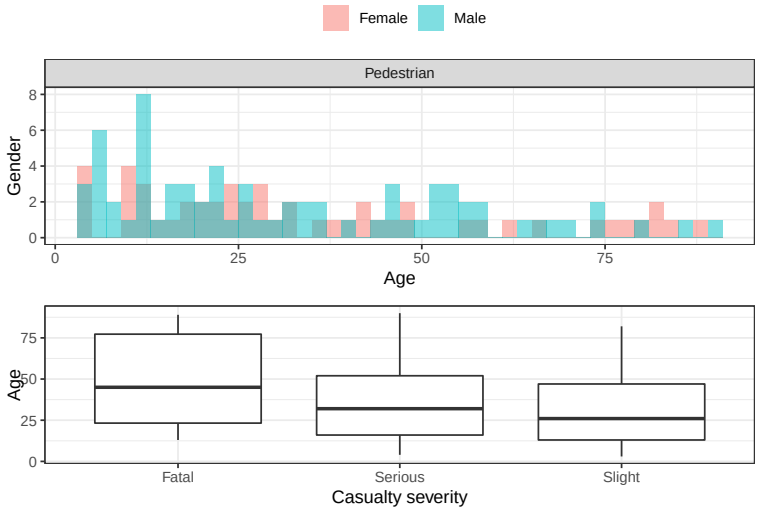

Figure 7: Comparing the age of casualty with gender

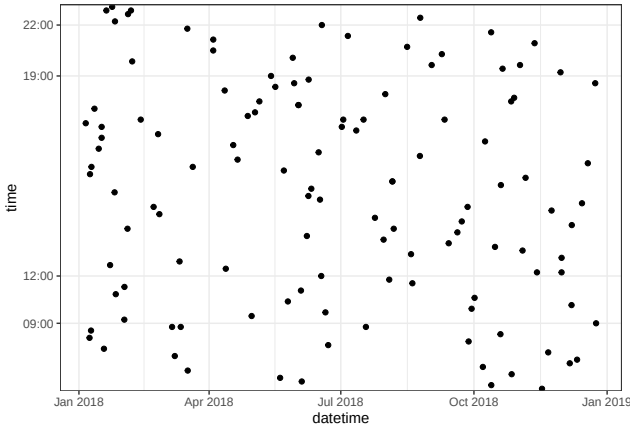

Figure 8: Time of the day for most casualty compared to different period of the year

3.2.3 Day of the week

Now, based on day of week, we will analyze when our pedestrians more prone to accidents. IN figure 10,As we see the accidents are highest at start of week, specifically on Tuesdays during 3-4pm followed by Monday and Friday around same time. Least accidents happen early mornings on any working day and Thursday and Wednesday in general see less accidents then rest of the weekdays. Fridays have less peaks but in general more casualties happen on Fridays than any other day.Sunday night time has more casualties.We can understand from this graph that maybe on weekends people are more susceptible to accidents maybe drunk drivers. We used lubridate() function to extract the year, month and day from the date variable.

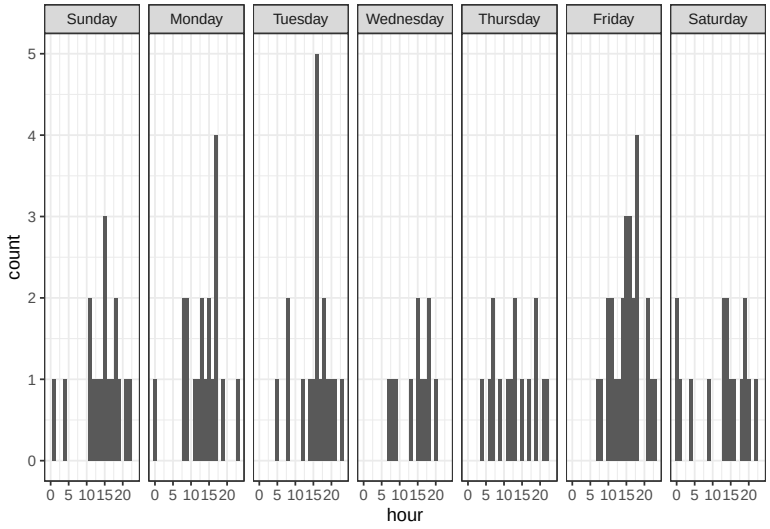

Figure 9: Pedestrians casualtiy by day of the week

3.2.4 Physical conditions

Since we analyzed our crash data in terms of time,day,age and sex we will now examine the physical conditions at the time of accidents and understand the severity of these accidents based on that.

In figure 11,Lightning conditions have hardly any impact on severity ratio,but as we can see by weather condition plots, high winds causes the most fatal accidents for pedestrians and serious injuries when it's raining due to lack of visibility. During winter season people are more vary and less severe accidents happens while crossing the roads, but wet and damp roads causes severe accidents along with snow. Rural areas have less traffic safety measures in places and people don't look around while crossing much, the reason why more fatal accidents happen around rural area than Urban. Urban have more less severe injuries than rural.

Discussion

Overall pedestrian casualty is more on weekends and start of week than rest of the days, weather conditions do play a role in signifying the visibility and hence lesser or more accidents.Most of the fatal injuries are for the age bracket 40-50 and more male than females. The urban area sees less fatal injuries than rural, given better roads and traffic rules. The walking percentage is governed by distance and area, in those area during particular seasons and time of day more fatality happens.

4 Conclusion

In conclusion, the level of walking in Leeds city is 34% which is pretty good compared to rest of the UK cities. The area and distance plays a key factor around for people to do walking trips. The pedestrians

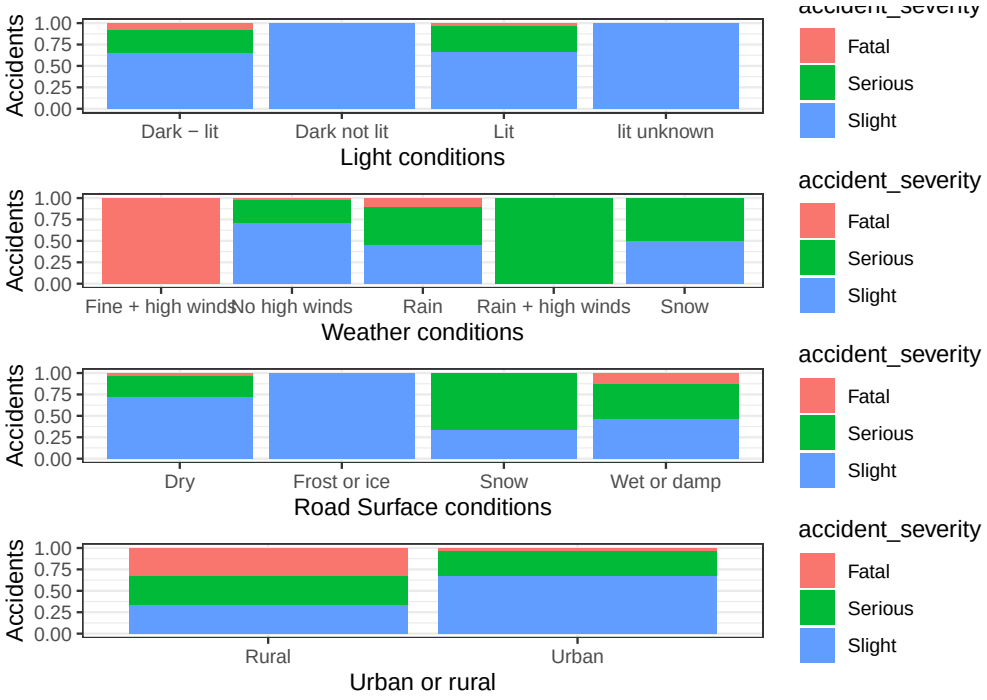

Figure 10: Physical conditions responsible for road safety

casualties also happens more in the areas where people prefer walking more, than other part of the city. Majority of pedestrian casualties happened in rural area even though less people prefer walking there, and more males were ill fated than women. Young children were victims of accidents but not fatal as compared to adults of age group 45-50. The weather and light conditions also play a factor when regarding safety of walkers. The weekends and early start of week see more casualties than rest of the week. Friday specially sees more accidents than rest of the week. If we have better traffic control in the place we would see lesser casualties. Better infrastructure development of the city and making commercial complexes across the city so that it's convenient for people to walk and get things, would help with increase in walking percentage. Government could run awareness programs on the benefits of walking for all age group and educate people on the different available options, parks, trails which could be used for the activity.

5 Drawbacks of the methods applied here

As we did have crashes data to compare the pct zones with to get more information on the crashes we were not able to analyze on level of walking in terms of demographics, like age gender and if they are commuting for work or leisure activity. We could have also used OSM data to find different paths and roads and analyze the PCT dataset with it, but even after multiple trial, OSM-converted sp object was not getting created with internet server issue.

We also assumed the crashes data available in stats19 is true and absolute, this might not hold true as this is the public record and might not have all accidents reported. Regarding the RMD file, while we are trying to maintain the file's consistency with graph it always fluctuates, even if we arrange it.

References

1.Docs.ropensci.org. 2021. Introduction to R for road safety: an introduction to R and practical exercises. [online] Available at: https://docs.ropensci.org/stats19/articles/stats19-training.html [Accessed 27 May 2021].

2.Robin Lovelace, J., 2021. Chapter 12 Transportation | Geocomputation with R. [online] Geocompr.robinlovelace.net. Available at: https://geocompr.robinlovelace.net/transport.html [Accessed 27 May 2021].

3.Blog.ukdataservice.ac.uk. 2021. How can we calculate levels of deprivation or poverty in the UK? (part 1) – Data Impact Blog. [online] Available at: http://blog.ukdataservice.ac.uk/deprived-or-live-in-poverty-1/ [Accessed 27 May 2021].

4.Living Streets. 2021. How walkable is Leeds?. [online] Available at: https://www.livingstreets.org.uk/get-involved/the-uks-top-walking-cities/how-walkable-is-leeds [Accessed 27 May 2021].

5.Travelindependent.org.uk. 2021. Road Casualties: Leeds Highway Authority Area. [online] Available at: http://www.travelindependent.org.uk/area_049.html [Accessed 27 May 2021].

6.nhs.uk. 2021. Walking for health. [online] Available at: https://www.nhs.uk/live-well/exercise/walking-for-health/ [Accessed 27 May 2021].

7.file:///Users/eshanashrivastava/Downloads/stats19-example%202/tds-coursework-example-reproducible.html

8.R Lovelace, L., 2021. Introducing stats19. [online] Cran.r-project.org. Available at: https://cran.r-project.org/web/packages/stats19/vignettes/stats19.html [Accessed 27 May 2021].

9.Pedestriansafety.org.uk. 2021. Road casualty analysis. [online] Available at: http://www.pedestriansafety.org.uk/stats.php [Accessed 27 May 2021].

10.Cran.r-project.org. 2021. Cycling potential in UK cities. [online] Available at: https://cran.r-project.org/web/packages/pct/vignettes/cycling-potential-uk.html [Accessed 27 May 2021].

11.Crashstats.nhtsa.dot.gov. 2021. [online] Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811748 [Accessed 27 May 2021].

12.Rdocumentation.org. 2021. pct package - RDocumentation. [online] Available at: https://www.rdocumentation.org/packages/pct/versions/0.2.2 [Accessed 27 May 2021].

13.Rstudio.github.io. 2021. Leaflet for R - Shapes. [online] Available at: https://rstudio.github.io/leaflet/shapes.html [Accessed 27 May 2021].