The Accidental Time Interval

**Group 127**

|  |
| --- |
| Ashish Narenkumar Dave |
| Dhruv Kalia |
| Girishwaran Kalidass |
| Prashant Nayana Jakkampudi |
| Tanwi Nandkeolyar |

Contents

[Abstract: 3](#_Toc61037283)

[I. Introduction 3](#_Toc61037284)

[II. Background 3](#_Toc61037285)

[III. Research design 4](#_Toc61037286)

[A. Research Objective and Research Question 4](#_Toc61037287)

[B. Understanding the data 4](#_Toc61037288)

[C. Data cleansing 5](#_Toc61037289)

[D. Data derivation 5](#_Toc61037290)

[E. Quantitative Analysis 5](#_Toc61037291)

[IV. Data Visualizations 6](#_Toc61037292)

[V. Result 8](#_Toc61037293)

[VI. Conclusion 8](#_Toc61037294)

[VII. Future scope 8](#_Toc61037295)

[VIII. References 9](#_Toc61037296)

# Abstract:

Car accidents are one of the most common causes of preventable deaths. Data on various attributes such as time, location, and other circumstances affecting the accidents can provide insights to frame policies, regulations, and necessary amendments in the traffic rules and regulations to prevent possible car accidents. This research is a statistical analysis of a publicly available dataset of 3.5 million car accidents that occurred in the United States over the past four years. We have categorized accidents according to their time of occurrence. The amount of traffic on roads can vary between a weekday and weekend. Thus, we try to identify specific time intervals with a higher frequency of car accidents on a weekday and a weekend. According to our findings, there is more number of accidents occurring in the morning as compared to other day and night times. However, after performing analysis of accidents occurring on weekdays and weekends separately, 7-9 am on a weekday and 11-1 pm on weekends are the most dangerous time intervals in terms of the frequency of car accidents.

# Introduction

As per a recent study done by World Health Organisation (WHO), 1.2 million people died due to road traffic accidents in a year mainly affecting younger men within 15-29 years age group (Ladeira *et al.*, 2017). The number of road traffic accidents have decreased significantly in high-income countries such as the US, UK and Germany (Ernstberger *et al.*, 2015). Information on the diurnal variations in accidents according to their numbers and severity can help make necessary changes in policies to further prevent mortality and disability due to road traffic accidents (Ernstberger *et al.*, 2015).

This study aims to identify time intervals in which most of the car accidents have occurred in the US. We are using a dataset available for academic, research and non-commercial purposes having data of car accidents which happened in the US. (Moosavi, Samavatian, Parthasarathy and Ramnath, 2019; Moosavi, Samavatian, Parthasarathy, Teodorescu, *et al.*, 2019).

# Background

Globally, extensive research has been conducted to assess the burden of accidents (proportions of accidents, mortality rate) to identify the factors affecting them. A Swedish study on the effect of traffic density on accidents has revealed that there are higher odds of fatal accidents occurring during early morning time (0400 hrs) and the night (Åkerstedt, Kecklund and Hörte, 2001). They followed an approach of breaking down a day into 4-hour intervals to conclude with the frequency of accidents in between them. Similar relevant studies can bring important changes in traffic rules and regulations. Therefore, country-wise detailed analysis of traffic accidents data is required to frame country-level traffic policies.

A project has been done under Udacity’s Data Scientists Nanodegree Program to analyse UK traffic Accidents from a dataset of the past 10 years (*rawanm/DataScientistNanodegree*, no date). Heatmaps have been plotted to visualize accidents occurring on a per hour and per day basis. They found out that Morning and evening rush-hours account for more number of accidents in a day.

# Research design

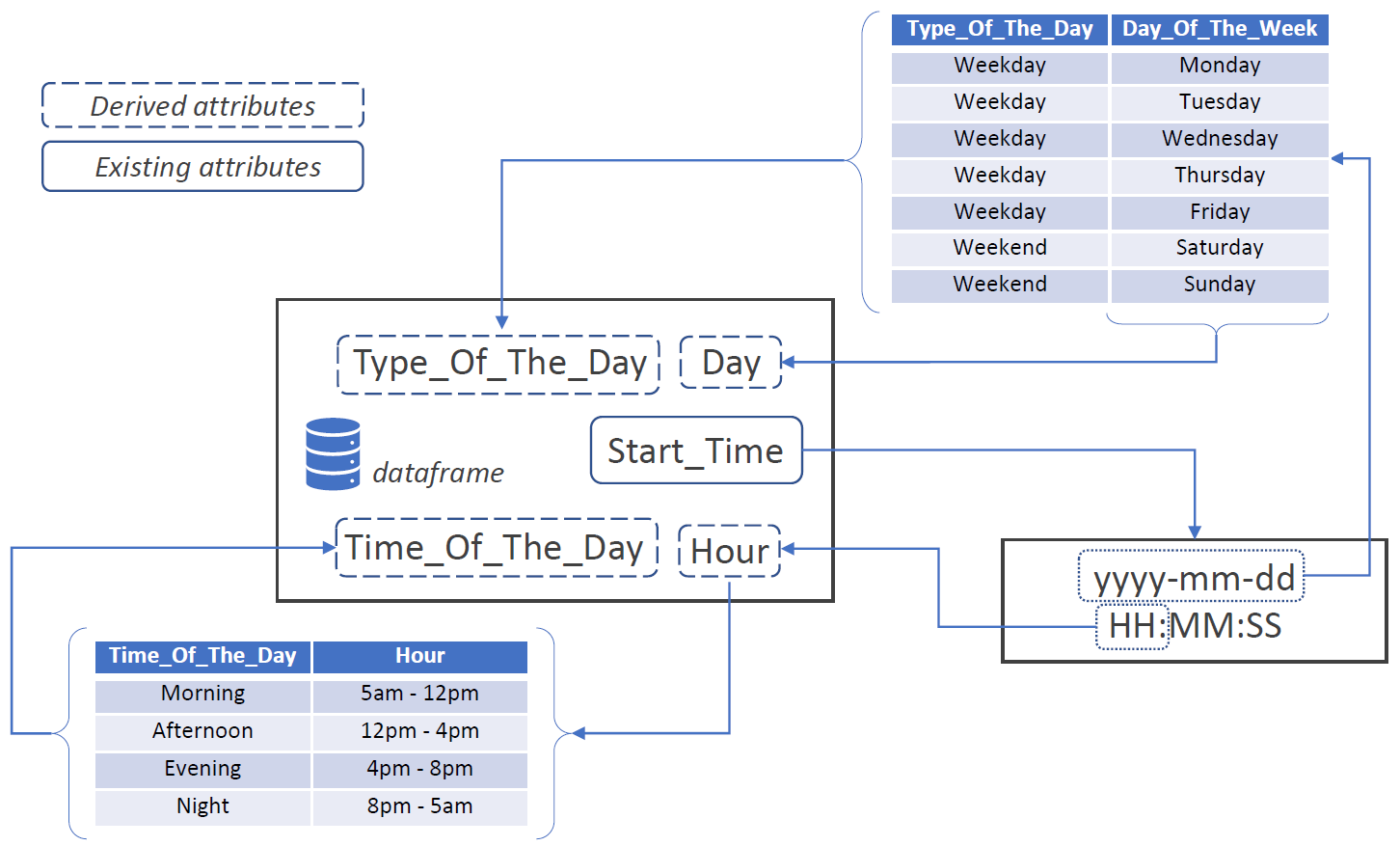


Figure 1 Using an existing column to derive dynamic attributes

## Research Objective and Research Question

In this research, we analyze if an equal number of car accidents occur at different times of the day (i.e., Morning, Afternoon, Evening, or Night) in the US. It would be a sheer coincidence if that did happen. Thus, our objective is to check if there exists a significant difference between the number of accidents occurring at different times of the day. Throughout this study, we try to find evidence to answer our research question:

**RQ**: *Is there a difference in proportions of accidents* *occurring at different times of the day (i.e., Morning/Afternoon/Evening/Night)?*

Our hypotheses for the above research question is mentioned in Table 1. By comparing the proportions of accidents occurring at different times of the day, we plan to identify a specific time interval in which there have been a greater number of accidents.

## Understanding the data

1. *Source:* We have chosen a country-wide car accident dataset available to everyone on Kaggle for non-commercial, research, and academic applications (Moosavi, Samavatian, Parthasarathy and Ramnath, 2019; Moosavi, Samavatian, Parthasarathy, Teodorescu, *et al.*, 2019). The accidents covered in this dataset are from 49 states of the USA between February 8, 2016, to July 1, 2020 (*US Accidents (3.5 million records)*, no date). This dataset was collected in real-time from different Traffic APIs.
2. *Key characteristics:* The dataset comprises of several attributes ranging from severity, location, time of the accident, presence of an airport, railway station, junction or crossing near the precise location of the accident.
3. *Choosing attributes relevant to RQ:* To find the count of accidents, we made use of the attribute: ID (a unique identifier for each accident). The dataset did not have any attribute showcasing the time of the day. Thus, we derived this value from the Start\_Time column.

## Data cleansing

The dataset has 3.5 million accidents with a unique ID and a Start\_Time. There are no duplicate or NA values in the two columns being used, i.e., ID and Start\_Time. All other attributes are dynamically created from these columns. Thus, no cleaning process was required for this research.

## Data derivation

Below is the process of deriving several columns from Start\_Time column. This has also been depicted in Figure 1.

1. *Process of deriving* Hour*:* Start\_Time column in our dataset had a value in the format of yyyy-mm-dd HH:MM:SS. The hour (HH) value was extracted from this column and stored in a new column, named Hour.
2. *Process of deriving* Time\_Of\_The\_Day*:* We have created a new attribute with the name, Time\_Of\_The\_Day. This can hold values like Morning, Afternoon, Evening, or Night. This denotes the time of the day in which the accident occurred. This column was populated based on the value of the previously calculated Hour column. Figure 1 shows how the hours are distributed with the time of the day.
3. *Process of deriving* Day\_Of\_The\_Week*:* In this new column, we have the days on which the accident took place. As always, we are utilizing the same Start\_Time column to get the date and thereafter, we are converting it to the day of the week.
4. *Process of deriving* Type\_Of\_The\_Day*:* This column can have two values, a weekday or a weekend, denoting on which type of the day the accident took place.

## Quantitative Analysis

At first, to identify any difference in proportions of accidents occurring at different times of the day, the χ2 value for H10 was calculated. Thereafter, the p-value was evaluated from the χ2 value. H10 can be rejected if its p-value does not fall in the range of critical values. Since a two-tailed test has been performed with a degree of freedom as 3, the critical value ranges from 0.001 to 5.024 for a significant value of 0.05.

|  |  |  |
| --- | --- | --- |
| Hypotheses | | Variables |
| H10 | There is no difference in the proportions of accidents occurring at different times of the day. | Proportion of occurrences of accidents  Times of the day (Morning, Afternoon, Evening, Night) |
| H20 | There is no difference in proportions of accidents occurring in different time intervals on a **weekday**. | Proportion of occurrences of accidents  1-hour time intervals between 24 hours on weekdays |
| H30 | There is no difference in the proportions of accidents occurring in different time intervals on a **weekend**. | Proportion of occurrences of accidents  1-hour time intervals between 24 hours on weekends |

Table 1 Hypotheses and its variables

For H20 and H30, a similar test is performed, and the p-value is calculated but with varied variables. In H20, accidents which occurred on weekdays are analyzed to check if a similar number of accidents occur at different hours. In H30, instead of weekdays, the focus is on the accidents which happened on Saturdays and Sunday.

# Data Visualizations

In Figure 2, Time\_Of\_The\_Day is plotted on the x-axis whereas the proportion of the accidents that occurred during that time is plotted on the y-axis. According to the bar plot, most of the accidents have occurred during the morning time (5 am – 12 pm).

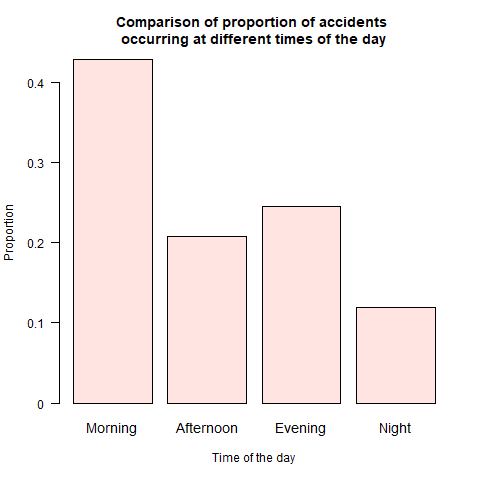


Figure 2 Proportion of accidents occurred during different times of the day

In Figure 3, we have segregated weekdays and weekends on the x-axis. This gives us a better understanding of the proportion of accidents occurring at different times of the day during a weekday and a weekend. As we can observe, the proportion of morning accidents that have occurred on a weekday is way more than on a weekend. However, on a weekend, the number of accidents occurring at different times of the day are almost the same proportion.

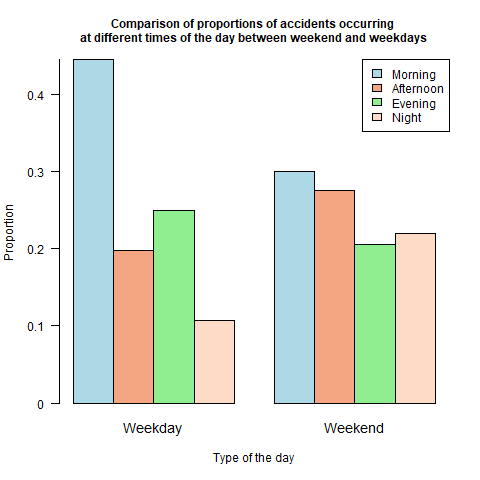


Figure 3 Proportion of accidents occurred during weekdays and weekends

As we can observe in Figure 3, the number of accidents on a weekend are uniformly distributed across different times of the day. To visualize this distribution in more depth, Figure 5 was plotted. Clearly, the plot is centrally skewed, and a greater number of accidents occur between 10 – 3 pm on a weekend. However, on a weekday, a greater number of accidents occur between 9 - 11 am (as shown in Figure 4).

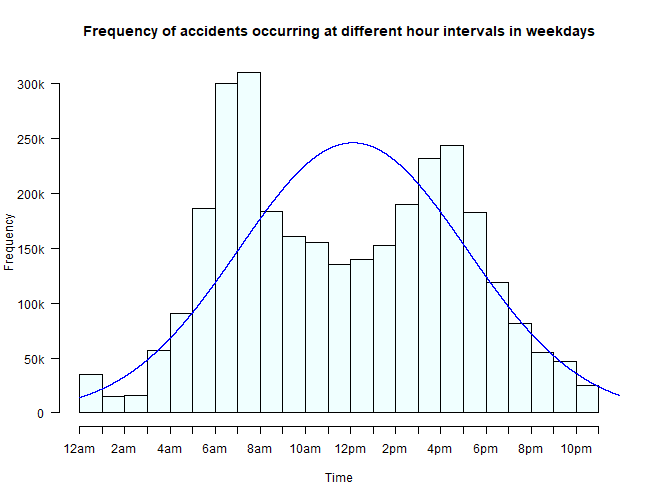


Figure 4 Frequency of accidents occurring at different time interval on a weekday

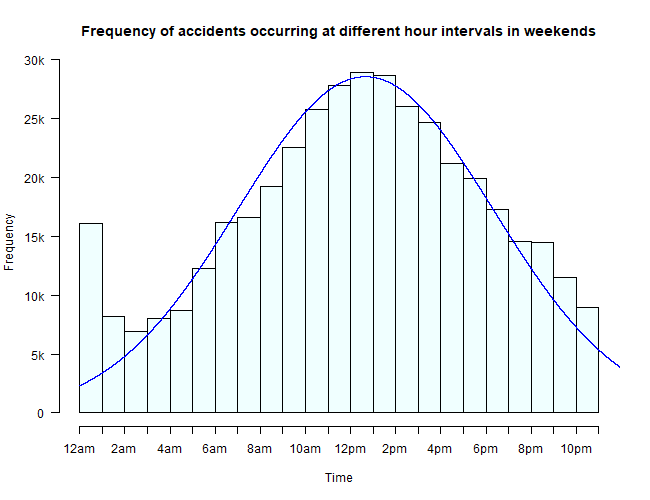


Figure 5 Frequency of accidents occurring at different time interval on a weekend

# Result

All the values of the result are shown in Table 2. The proportion of car accidents were significantly higher (p-value < 0.05) during the morning as compared to other times of the day (i.e., afternoon, evening, and night). Thus, we can reject our null hypotheses (H10). This implies, that there exists a significant difference between the number of accidents occurring at different times of the day.

The results for χ2 test for H20 and H30 are like H10. All the p-values are less than the significant level (0.05). With the result in Table 2, we can reject all three hypotheses made in Table 1. Thus, we can say, that there is strong evidence that there is a difference in the proportion of car accidents occurring at different time-periods of the day irrespective of the day being a weekday or a weekend.

|  |  |  |  |
| --- | --- | --- | --- |
| Hypotheses | χ2 | degree of freedom | p-value |
| H10 | 953555 | 3 | < 2.2e-16 |
| H20 | 1331401 | 22 | < 2.2e-16 |
| H30 | 74305 | 22 | < 2.2e-16 |

Table 2 Result

# Conclusion

In this paper, we have analysed a data set on car accidents which have occurred in the United States of America. As per the findings, there exists a significant difference in the number of accidents that have occurred during different times of the day. To enhance our perspective and to form a strong conclusion, we segregated weekdays from weekends and introduced two new hypotheses.

It was observed that the higher proportion of accidents on weekdays occur between 7 am to 9 am. This could be due to the heavy traffic on roads or people being in a hurry to reach their work/school/college on time.

Our results reveal, there exists a difference in the proportion of accidents occurring at different times of the day on a weekend. As per Figure 5, the number of accidents between 10 – 3 pm is more when compared to any other time interval of the day on a weekend. This could be because people tend to start their day with a bit more ease and consequently, traffic on roads is spread throughout the day.

# Future scope

This study has some great potential that can yield significant results in predicting the high-risk zones where most of the accidents occur. As weather plays a crucial role while driving, it can affect the severity of an accident. Traffic authorities can use this data to prepare their emergency vehicles for future accidents occurring due to extreme climatic conditions.

Furthermore, this was the 3rd version of the dataset that was published in July 2020. Next version shall be released for public soon. It will have data from July – December 2020. With the increase and latest update in the number of accidents, the accuracy and estimations of predictions and analysis should improve. We might also see some new studies done with this dataset in the coming future.

# References

1. Åkerstedt, T., Kecklund, G. and Hörte, L.-G. (2001) ‘Night Driving, Season, and the Risk of Highway Accidents’, *Sleep*, 24(4), pp. 401–406. doi: 10.1093/sleep/24.4.401.
2. Ernstberger, A. *et al.* (2015) ‘Decrease of morbidity in road traffic accidents in a high income country – an analysis of 24,405 accidents in a 21 year period’, *Injury*, 46, pp. S135–S143. doi: 10.1016/S0020-1383(15)30033-4.
3. Ladeira, R. M. *et al.* (2017) ‘Acidentes de transporte terrestre: estudo Carga Global de Doenças, Brasil e unidades federadas, 1990 e 2015’, *Revista Brasileira de Epidemiologia*, 20(suppl 1), pp. 157–170. doi: 10.1590/1980-5497201700050013.
4. Moosavi, S., Samavatian, M. H., Parthasarathy, S. and Ramnath, R. (2019) ‘A Countrywide Traffic Accident Dataset’, *arXiv:1906.05409 [cs]*. Available at: http://arxiv.org/abs/1906.05409 (Accessed: 7 January 2021).
5. Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., *et al.* (2019) ‘Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights’, in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. *SIGSPATIAL ’19: 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Chicago IL USA: ACM, pp. 33–42. doi: 10.1145/3347146.3359078.
6. *rawanm/DataScientistNanodegree* (no date) *GitHub*. Available at: https://github.com/rawanm/DataScientistNanodegree (Accessed: 8 January 2021).
7. *US Accidents (3.5 million records)* (no date). Available at: https://kaggle.com/sobhanmoosavi/us-accidents (Accessed: 8 January 2021).