**SAMUEL WAIRIUKO**

**SAS PROJECT REPORT**

**Barcelona Car Accidents 2017**

**Date 2020-06-06**

TABLE OF CONTENTS

Executive summary…………………………………………………………………………………………………………………….………….2

Introduction .................................................................................................................................................2

Properties of the dataset ..............................................................................................................................2

Objectives………………………………………………………………………………………………………………………………………….……2

Conceptual Framework .......................... .....................................................................................................3

Hypotheses ..................................................................................................................................................3

Methodology ................................................................................................................................................4

PROC Statements Used .................................................................................................................................4

Data Preprocessing .......................................................................................................................................5

Outlier Detection ..........................................................................................................................................5

Descriptive Analysis ......................................................................................................................................5

Univariate Analysis ....................................................................................................................................6-9

Feature Engineering………………………………………………………………………………………………………………………………10

Bivariate Analysis ..................................................................................................................................11-12

Chi-square Analysis ....................................................................................................................................13

Scatterplot..................................................................................................................................................14

Inferential Analysis .....................................................................................................................................15

Correlation Analysis……………………………………………………………………………………………………………….………………16

One-Sample T-test ......................................................................................................................................17

Linear Regression .......................................................................................................................................18

Summary of Results ....................................................................................................................................19

Recommendations .....................................................................................................................................19

SAS Scripts ..................................................................................................................................................20

**Executive summary**

Speeding, careless driving, driving under influence of alcohol and poor weather conditions are major causes of accidents. Living in densely populated areas means high number of people driving to and from the workplace thus congested traffic. In Barcelona Spain the total number of reported accidents in 2017 as the dataset provided by the lecturer was 10339. This analysis aims at finding out the locations and time when most accidents happen and also provides relevant insights to the public safety of Barcelona to put more safely precautions thus reducing the crashes.

**Introduction**

According to info Barcelona on road safety, the driver error, lack of concentration, not respecting safety distances and driving at inappropriate speed are major causes of road accidents. In addition, pedestrians disobeying the traffic lights, not using pedestrian crossings and walking on the roads are also main causes of the road accidents. In 2019 there were 9251 accidents which was around 25 accidents a day.

**Properties of the dataset**

Barcelona Car Accidents 2017 dataset had 10339 observations and 15 variables which includes; id, district name, neighborhood, street, month, day, weekday, part of the day, hour, mild Injuries, serious injuries, victims, vehicles involved, longitude and latitude. My target variable is Victims(deaths).

**Objectives**

The following were the objectives of the analysis:

* To find out the districts and streets with most victims in Barcelona.
* To know which part of the day, weekday, hour and month with the most accidents.
* To know the major causes of these accidents.
* To come up with insights that can be implemented to reduce the number of accidents.

**Conceptual Framework**

I divided the dataset in three different categories: location (neighborhood, district, street, longitude and latitude), time (month, weekday, day, part of the day, hour) and crash statistics (mild Injuries, serious injuries, victims, vehicles involved).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | DEPENDENT VARABLE(y)   |  | | --- | |  | |  |  |  |
|  | INDEPENDENT VARIABLES(X) | | | | | |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| LONGTITUDE  LATITUDE  DISTRICTNAME  NEIGHBOUR HOOD  STREET  MONTH  DAY  HOUR  WEEKDAY  PART OF THE DAY  VEHICLES INVOLVED  SERIOUS INJURIES  MILD INJURIES  VICTIMS  Time of Crash  Crash Statistics  Target(Y)  Location Coordinates  Location   |  | | --- | |  | |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

**Hypothesis**

* Which district, neighbourhood, street had the highest vehicle accidents?
* Which hour, part of the day, day, weekday, month do most accidents happen in Barcelona?
* Is there any association between weekday and part of the day?
* Is there any relationship between the number of vehicles involved and the Victims?

**Methodology**

* Used SAS 9.4 Statistical software for the Analysis.
* Performed data transformation through grouping the victim’s column in to five groups.
* Generate new features “Death “to track the locations with the highest number of deaths.

**PROC Statements Used**

Used the following PROC Statements used to analyze my findings:

* Proc Import: Used it to import the csv file.
* Proc Contents: Used it to see the contents in the dataset.
* Proc Print: Used it to display values for all observations and variables from the dataset.
* Proc Sort: Used it to sort data in descending order from highest to the lowest.
* Proc Means: Used to calculate basic statistics like five number summaries.
* Proc Freq: Used it to produce contingency tables with the summary totals of variables
* Proc Format: Used it for mapping data values into data labels.
* Proc SQL: Used it to create new variables.
* Proc Univariate: Used it in the discovery of outliers
* Proc SGPLOT: Used it to produce graphs, scatter plots and charts
* Proc Reg: Used it to fit the linear regression model.
* Proc Anova: used it to measure the differences in the means of variables.
* Proc Corr: used it to conduct Pearson correlation between two continuous variables.
* Proc Gchart: used bar charts to graphically represent my findings.
* Proc Summary: used it to output the counts.
* Proc Sgmap: to create maps.

**Data Preprocessing**

This is the process of cleaning the data and transforming it in to understandable format.

During the preprocessing I checked the missing values for the numeric continuous variables using five number summaries, then checked character missing values and outliers. There were 27 unknowns equal to 0.02% of the total observations, I dropped the id column, missing values and outliers.

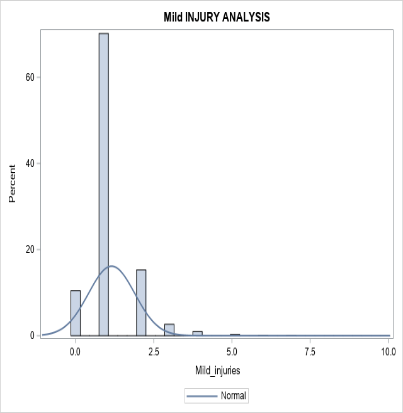
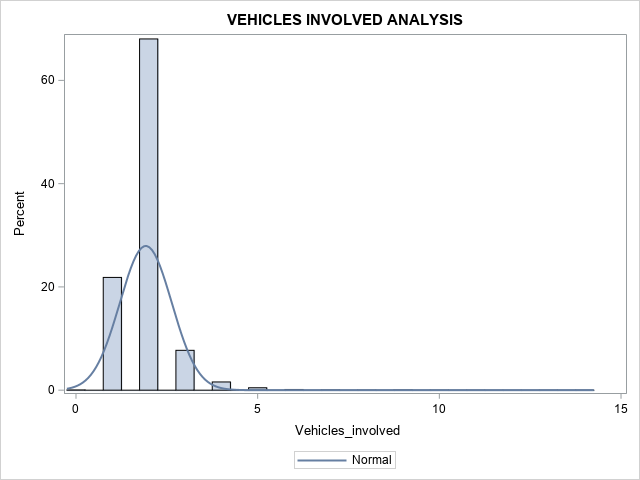
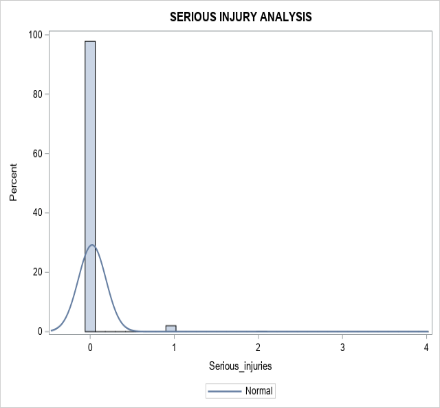
**Descriptive Analysis**

This is the descriptive statistics summarizing the features of the dataset.

The proc means procedure (checking missing values on numeric variables)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | N Miss | Minimum | Maximum | Mean | Median | Std Dev |
| Day Hour Mild\_injuries Serious\_injuries Victims Vehicles\_involved Longitude Latitude | 10339 10339 10339 10339 10339 10339 10339 10339 | 0 0 0 0 0 0 0 0 | 1.00 0.00 0.00 0.00 0.00 0.00 2.09 41.32 | 31.00 23.00 10.00 4.00 10.00 14.00 2.22 41.47 | 15.78 13.81 1.15 0.02 1.18 1.92 2.16 41.40 | 16.00 14.00 1.00 0.00 1.00 2.00 2.16 41.40 | 8.76 5.32 0.74 0.16 0.74 0.71 0.02 0.02 |

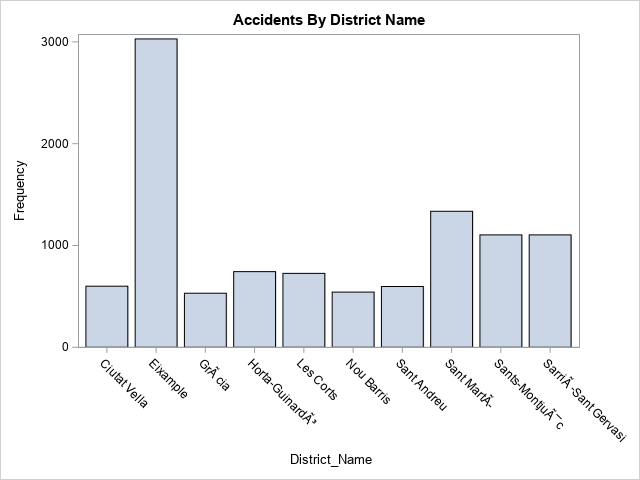
The maximum serious injuries were 4, maximum vehicles involved 14 and maximum victims 10.

The density curve of Mild injuries,Serious injuries and Vehicles involved shows normal distribution and data skewed towards the right. Possibility of Outliers.

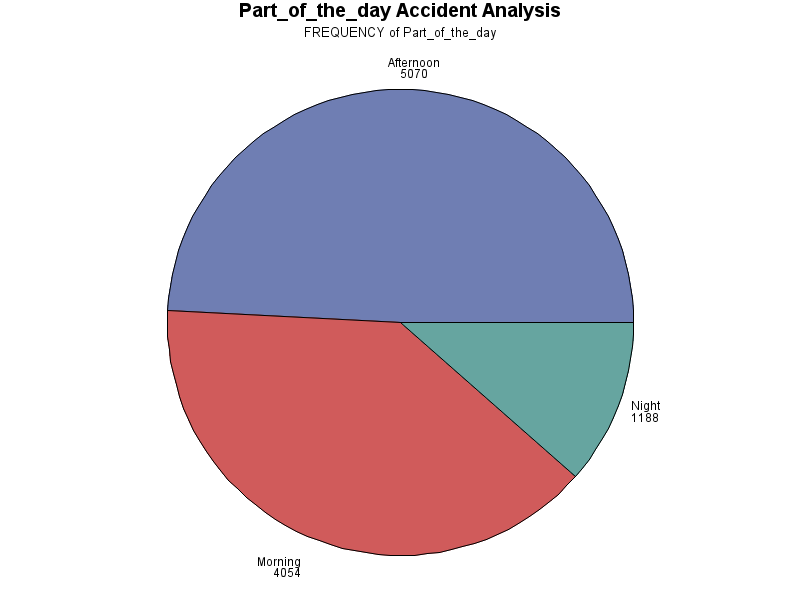
The outliers were deleted.

**Univariate analysis**

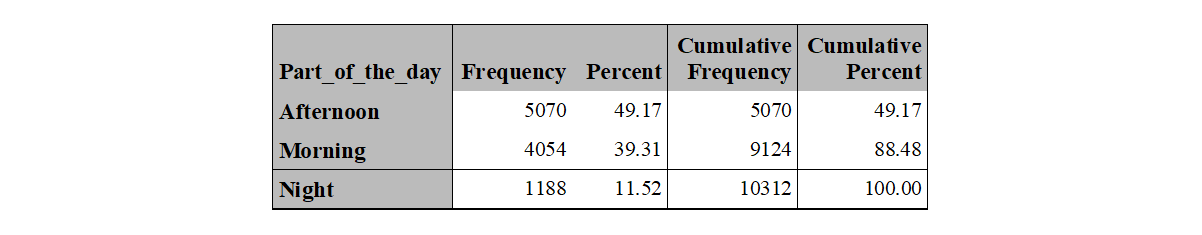
Univariate analysis shows the distribution of each variable in the dataset, to describe the data and find the patterns in it.

The district with the highest number of car accidents is Eixample with 29.37 %of the total accidents followed by Sant MartÃ­, Sants-MontjuÃ¯c and SarriÃ -Sant Gervasi as shown in the table below.

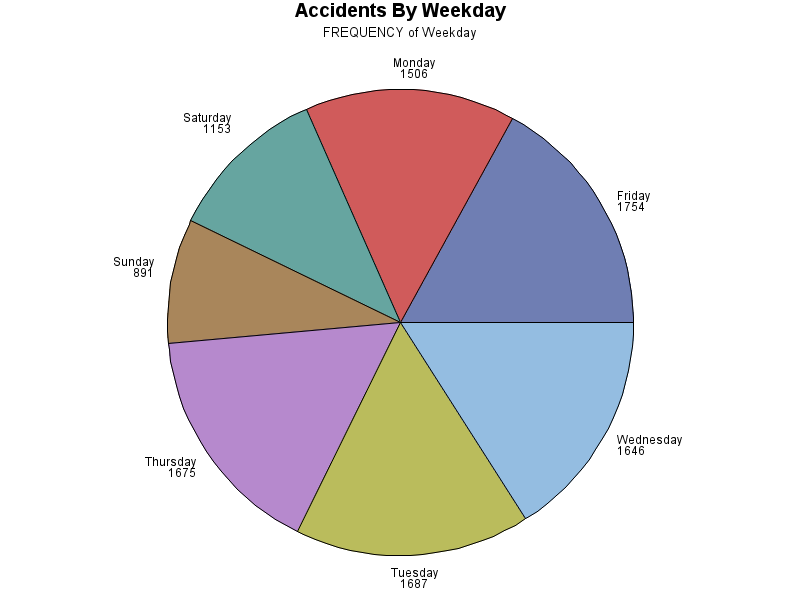
| **District\_Name** | **victims\_desc** |
| --- | --- |
| Eixample | 3562 |
| Sant MartÃ­ | 1645 |
| Sants-MontjuÃ¯c | 1329 |
| SarriÃ -Sant Gervasi | 1290 |
| Les Corts | 894 |
|  |  |



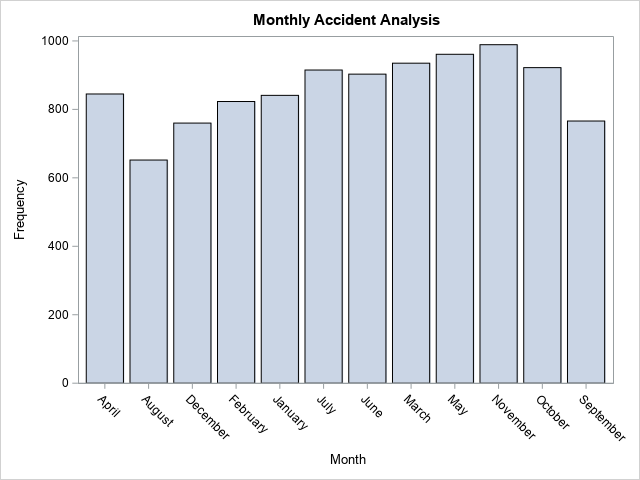
Afternoon had the highest number of accidents at 5070, followed by morning with 4054 and night with 1188 accidents.



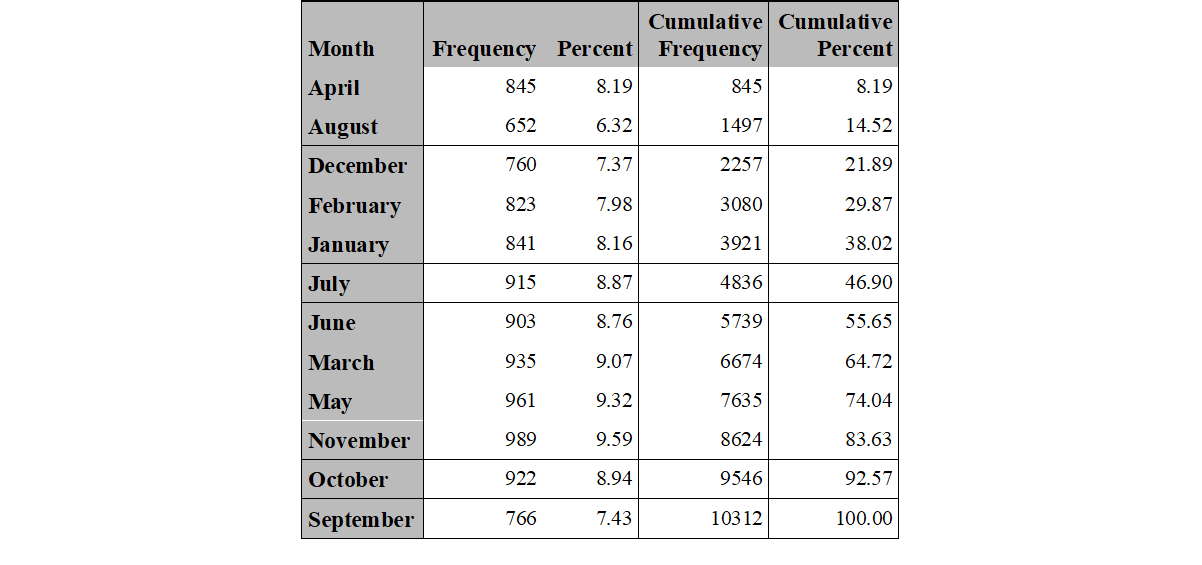
Afternoon was leading with number of accidents at 49% followed by Morning 39% and fewer accidents on Sunday at 11%.



Friday recorded the highest number of accidents with 1754 accidents. Sunday had the lowest number at 891 accidents.



November had the greatest number of car accidents with 989(9.59%). The lowest was August with 652(6.32%) accidents.

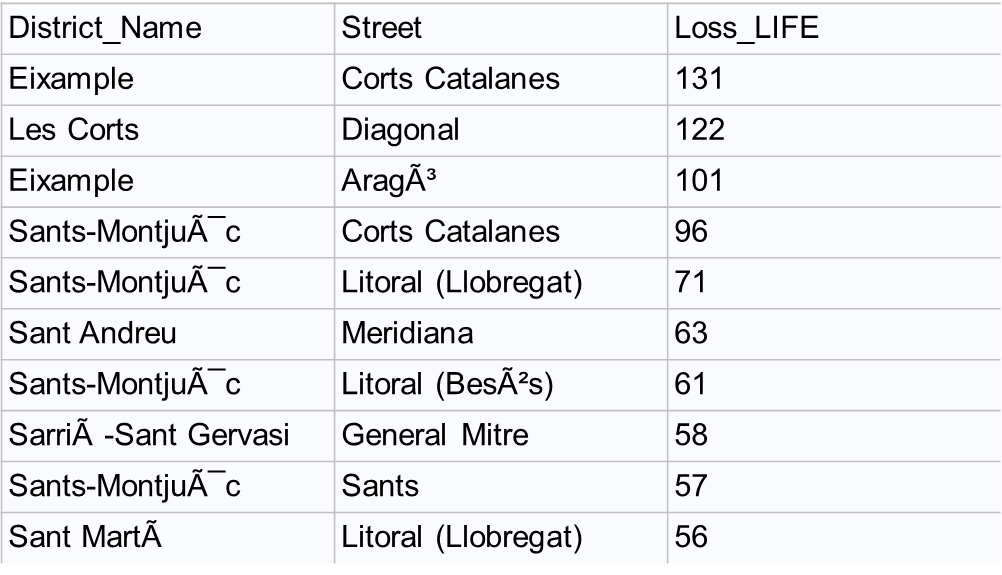


**Feature Engineering**

This is the process of using the domain knowledge to construct new features from the existing features. In this case I created “DEATHS “column to know the top 10 neighbourhoods with the highest number of victims.

|  |  |  |
| --- | --- | --- |
| District\_Name | Neighborhood Name | DEATHS |
| Eixample | la Dreta de l'Eixample | 1368 |
| Eixample | l'Antiga Esquerra de l'Eixample | 680 |
| Eixample | la Sagrada FamÃ­lia | 457 |
| Eixample | la Nova Esquerra de l'Eixample | 454 |
| SarriÃ -Sant Gervasi | Sant Gervasi – Galvany | 446 |
| Les Corts | les Corts | 385 |
| Eixample | el Fort Pienc | 319 |
| Sants-MontjuÃ¯c | la Marina del Prat Vermell | 304 |
| Sants-MontjuÃ¯c | el Poble-sec | 291 |
| Eixample | Sant Antoni | 284 |

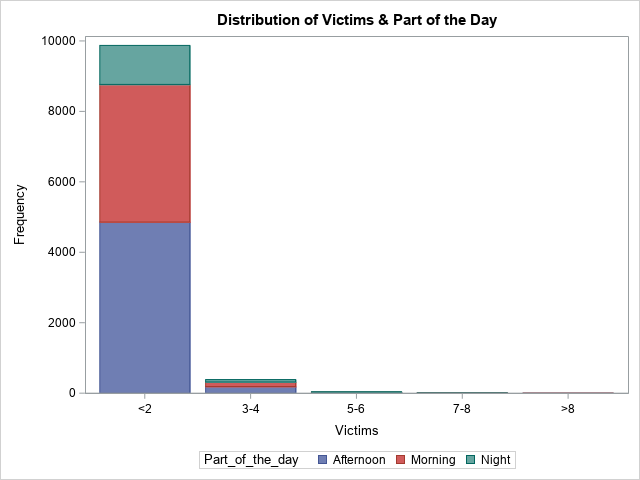
The neighbourhood with the greatest number of car accidents is la Dreta de l’Eixample. This neighborhood is in the most central of the Barcelona with the highest concentration of modern structure. The people living in the area are high class.



Corts Catalanes is the leading street in Eixample district with the highest number of victims.

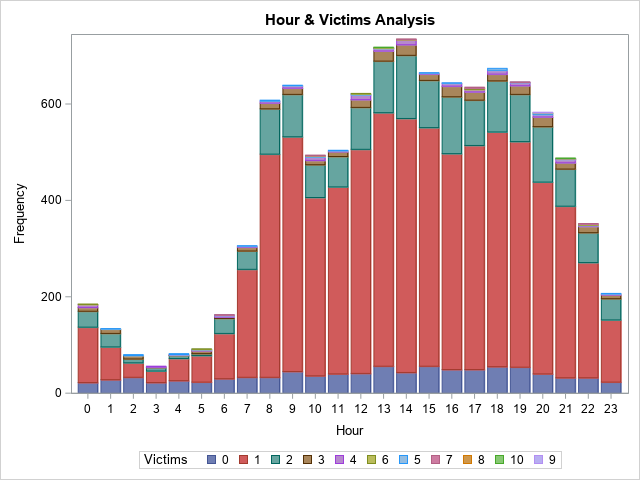
**Bivariate Analysis**

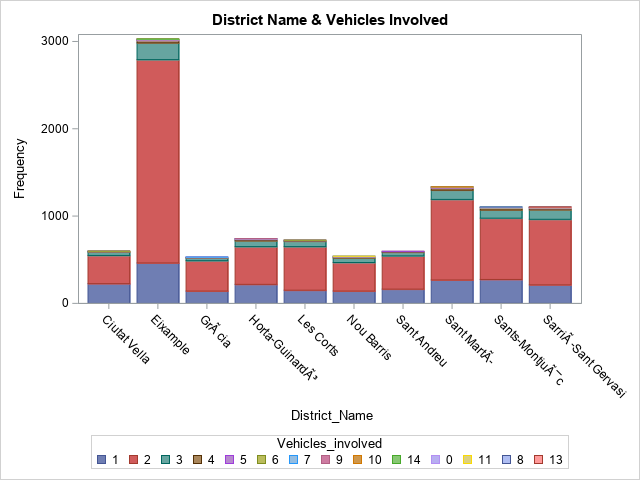
This is statistical analysis where two variables are observed dependent and independent and compare both variables to each other.



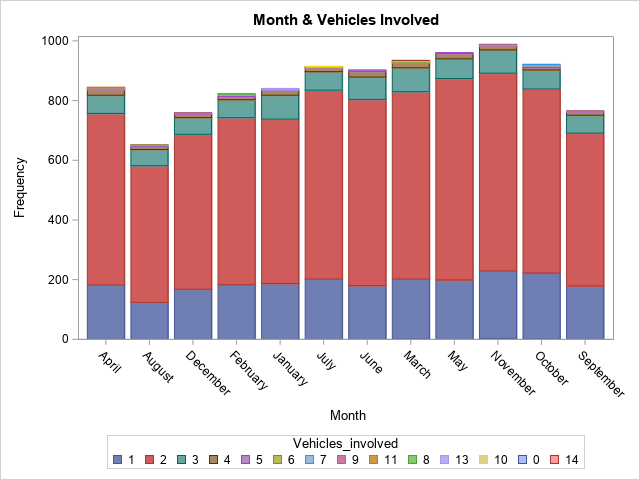
On average less than or equal 2 people lost their lives, most of the accidents happened in the afternoon followed by morning and few at night. Afternoon had the highest number of victims with 49% of the total accidents, followed by morning with 39.31% and night with 11.52% as shown in the above graph.

Most accidents happen between 8am-9am,13pm-14pm and 15pm-19pm with average of 1 victim as shown below.





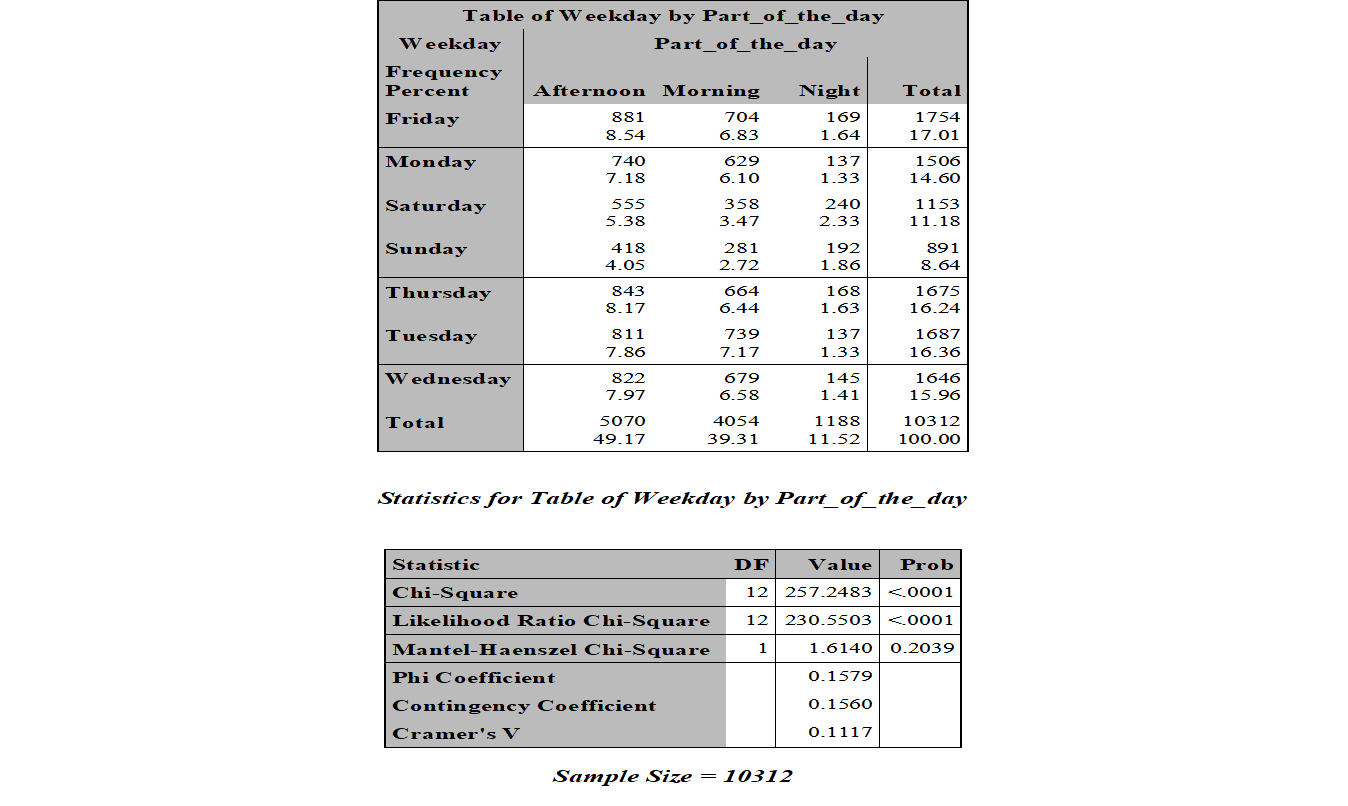
Eixample District recorded the highest number of accidents. Most of the accidents involved 2 vehicles.



Majority of the accidents involved 2 vehicles and highest number recorded in November.

**Chi-Square Statistics**

This is a test to measure how expectations compare to actual observed data.



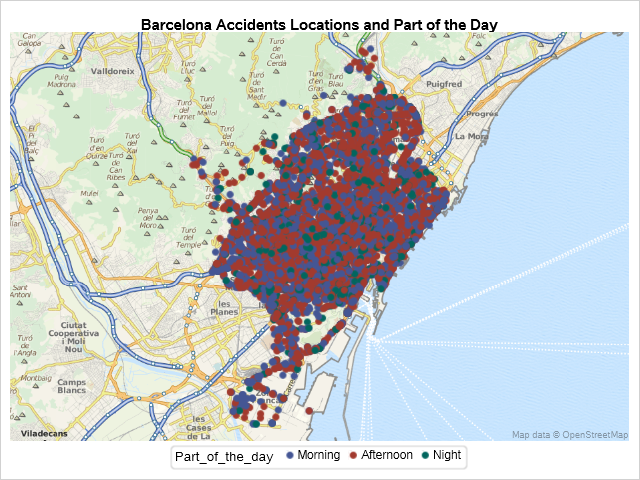
Most accidents happened Friday afternoon at 8.54% and lowest on Tuesday and Monday night both days at 1.33%. The chi-square shows the variables are not associated because phi coefficient is 0.15 close to zero.

**Scatter plot**

Used cartesian coordinates to graph the specific locations in Barcelona where accidents happened.

**Sgmap**

The density scatter plot for the latitude and longitude vs part of the day shows a liner tread.



**Inferential Analysis**

**Univariate Analysis**

Proc univariate procedure shows P-value (<.0001) meaning No significance difference in the means of victims.

****

**Correlation Analysis**

The correlation measures the relationship between two variables.

VICTIMS AND SERIOUS INJURIES

Pearson correlation between victims and serious injuries is 0.07, which is not significant with a p value of <.0001 indicating a weak relationship between the two variables.

VICTIMS AND MILD INJURIES

Pearson correlation between victims and mild injuries is 0.97, which is significant with a p value of <0. 0001.this indicates a strong linear positive relationship between the two variables.



**T-test Statistics**

Students t distribution is used to determine if there is significant between the means of two groups which may be related in certain features. I used one sample test to compare sample mean with the mean of population.



The distribution of victims is normal with 95% confidence interval between the sample mean and population mean.

**Linear Regression**

Conducted a linear regression model and R Square was 0.0247 this shows no linear relationship between the vehicles involved and the number of victims.



**Summary of Results**

* Eixample district have the highest number of accidents.
* Majority of la Dreta de l'Eixample neighbourhood belong to upper class of Barcelona with large selection of vacation destinations and high population.
* Most of the accidents happen in the afternoon and in the morning.
* Friday has the highest number of accidents and Sunday the lowest.
* The month with the highest number of accidents is November and the one with lowest is August.
* No linear relationship between the number of vehicles involved and Victims.

**Recommendations**

* The government of Barcelona should limit the speed of motorists.
* Road expansion.
* Install more speed cameras and traffic lights.
* The drivers to keep an eye on the road and avoid distractions.
* Implementing other prediction models for better results.

LIBNAME CAR"C:\Users\User\Desktop\CAR";

**PROC** **IMPORT** OUT= CAR.PROJECT

DATAFILE= "C:\Users\User\Desktop\CAR\accidents\_2017.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

GUESSINGROWS=**1000**;

**RUN**;

**PROC** **CONTENTS** DATA=CAR.PROJECT;

**RUN**;

\*PROC CONTENTS VARIABLES;

\*Id District\_Name Neighborhood\_Name Street Weekday Month

Day Hour Part\_of\_the\_day Mild\_injuries

Serious\_injuries Victims Vehicles\_involved Longitude Latitude

\*descriptive statistics for continuous data;

**PROC** **PRINT** DATA=CAR.PROJECT;

**RUN**;

\*CHECKING FOR MISSING VALUES NUMERIC VARIABLES ;

\*No missing values were found on continuous variables;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **MEANS** DATA =CAR.PROJECT MAXDEC= **2** N NMISS MIN MAX MEAN MEDIAN STD;

VAR Day Hour Mild\_injuries Serious\_injuries Victims Vehicles\_involved Longitude Latitude;

TITLE"MISSING VALUES CHECK";

**RUN**;

ODS RTF CLOSE;

\*CHECKING FOR MISSING IN CHARACTER VARIABLES;

\*No missing values;

\*PROC FORMAT;

\*VALUE $CHARS " " = "MISSING"

OTHER = "NOT MISSING";

\*RUN;

**PROC** **FREQ** DATA = CAR.PROJECT;

TABLE Neighborhood\_Name/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

**PROC** **FREQ** DATA = CAR.PROJECT;

TABLE Street/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

**PROC** **FREQ** DATA = CAR.PROJECT;

TABLE Weekday/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

**PROC** **FREQ** DATA = CAR.PROJECT;

TABLE Month/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

**PROC** **FREQ** DATA = CAR.PROJECT;

TABLE Part\_of\_the\_day/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

\*27 unknown districts;

**PROC** **FREQ** DATA = CAR.PROJECT;

TABLE District\_Name/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

\*treating missing values;

**PROC** **CONTENTS** DATA=CAR.PROJECT;

**RUN**;

**DATA** CAR.PROJECT1;

SET CAR.PROJECT;

IF District\_Name eq"Unknown" THEN DELETE;

**RUN**;

**PROC** **CONTENTS** DATA=CAR.PROJECT1;

**RUN**;

**DATA** CAR.PROJECT2;

SET CAR.PROJECT1;

LOWER\_LIMIT = Q1 - (**3**\*IQR);

UPPER\_LIMIT = Q3 + (**3**\*IQR);

DROP \_TYPE\_ \_FREQ\_;

**RUN**;

\*PRODUCING STATISTICS WITH PROC MEANS;

\*FIVE NUMBER SUMMARY;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **MEANS** DATA =CAR.PROJECT2 MAXDEC= **2** N MIN MAX MEAN Median q1 q3 ;

VAR Latitude Longitude Mild\_injuries Serious\_injuries Vehicles\_involved Victims;

**RUN**;

ODS RTF CLOSE;

**PROC** **MEANS** DATA =CAR.PROJECT2 MAXDEC= **2** N Median q1 q3 ;

VAR Latitude Longitude Mild\_injuries Serious\_injuries Vehicles\_involved Victims;

**RUN**;

**PROC** **MEANS** DATA =CAR.PROJECT2 MAXDEC= **2** N MIN MEAN MAX Median q1 q3;

VAR Latitude Longitude Mild\_injuries Serious\_injuries Vehicles\_involved Victims;

**RUN**;

**PROC** **MEANS** DATA = CAR.PROJECT2 MAXDEC=**2** ;

VAR Latitude;

OUTPUT OUT=TEMP P25 =Q1 P75 = Q3 QRANGE = IQR;

**RUN**;

\*ACCIDENTS SUMMARY ;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**proc** **summary** data=CAR.PROJECT2;

var Mild\_injuries Serious\_injuries Vehicles\_involved Victims;

output out=sumdata;

TITLE"ACCIDENTS SUMMARY TABLE ";

**run**;

**PROC** **PRINT** DATA=sumdata;

**RUN**;

ODS RTF CLOSE;

\*UNIVARIATE ANALYSIS CONTINUOUS;

\*Examining distribution of the data;

**PROC** **SGPLOT** DATA = CAR.PROJECT2 ;

HISTOGRAM Latitude ;

TITLE"LATITUDE DISTRIBUTION ";

**RUN**;

**PROC** **SGPLOT** DATA = CAR.PROJECT2 ;

HISTOGRAM Longitude;

TITLE"LONGITUDE DISTRIBUTION ";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

HISTOGRAM Vehicles\_involved;

DENSITY Vehicles\_involved;

TITLE"VEHICLES INVOLVED ANALYSIS ";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

HISTOGRAM Serious\_injuries;

DENSITY Serious\_injuries;

TITLE"SERIOUS INJURY ANALYSIS ";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

HISTOGRAM Mild\_injuries;

DENSITY Mild\_injuries;

TITLE"Mild INJURY ANALYSIS ";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

HISTOGRAM Victims;

DENSITY Victims;

TITLE"Victims ANALYSIS ";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

HISTOGRAM Latitude;

DENSITY Latitude;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

HBOX Longitude;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBOX Latitude;

**RUN**;

\*CATEGORICAL COLUMNS;

\*UNIVARIATE ANALYSIS;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Hour;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR District\_Name;

TITLE "Accidents By District Name";

**RUN**;

**proc** **gchart** data=CAR.PROJECT2;

pie Weekday ;

TITLE "Accidents By Weekday";

**run**;

**proc** **gchart** data=CAR.PROJECT2;

pie Part\_of\_the\_day ;

TITLE "Part\_of\_the\_day Accident Analysis";

**run**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Month;

TITLE "Monthly Accident Analysis";

**RUN**;

**PROC** **SORT** DATA1= CAR.PROJECT2 OUT= bercelona\_Day\_SORTED;

BY Day;

**RUN**;

**PROC** **SGPLOT** DATA=bercelona\_Day\_SORTED;

HISTOGRAM Day;

DENSITY Day;

TITLE "Day Accident Analysis";

**RUN**;

**RUN**;**PROC** **SORT** DATA = CAR.PROJECT2 OUT= bercelona\_Day\_SORTED;

BY Day;

**RUN**;

**PROC** **SGPLOT** DATA=bercelona\_Day\_SORTED;

HISTOGRAM Day;

DENSITY Day;

**RUN**;

**PROC** **SORT** DATA = CAR.PROJECT2 OUT= bercelona\_Hour\_SORTED;

BY Hour;

**RUN**;

**PROC** **SGPLOT** DATA=bercelona\_Hour\_SORTED;

HISTOGRAM Hour;

DENSITY Hour;

**RUN**;

**proc** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Month ;

FORMAT Month SYEARgrp.;

**run**;

\*PROC SGPLOT BIVARIATE ANALYSIS;

**proc** **format**;

VALUE VGRP **0**-**2**="<2"

**3**-**4**="3-4"

**5**-**6**="5-6"

**7**-**8**="7-8"

**9**-**10**=">8";

TITLE"VICTIMS GROUPED";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Part\_of\_the\_day/GROUP=Victims;

TITLE"VICTIMS VS PART OF THE DAY";

FORMAT Victims VGRP.;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Hour/GROUP=Serious\_injuries;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Weekday/GROUP=Victims;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Weekday/GROUP=Vehicles\_involved;

TITLE"VEHICLES INVOLVED VS WEEKDAY";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Month/GROUP=Vehicles\_involved;

TITLE"Month & Vehicles Involved ";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Day/GROUP=Vehicles\_involved;

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR District\_Name/GROUP=Vehicles\_involved;

TITLE"District Name & Vehicles Involved";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR District\_Name/GROUP=Serious\_injuries;

**RUN**;

**proc** **sgplot** data=CAR.PROJECT2;

histogram Vehicles\_involved / group=Part\_of\_the\_day transparency=**0.5** scale=count;

density Vehicles\_involved / type=normal group=Part\_of\_the\_day;

keylegend / location=inside position=topright across=**1**;

**run**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Hour/GROUP=Victims;

TITLE "Hour & Victims Analysis";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT1;

VBAR Part\_of\_the\_day/GROUP=Victims;

TITLE"Part of the Day and Victims Analysis";

**RUN**;

\* CORRELATION STATISTICS;

\*PROC CORR;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **CORR** DATA=CAR.PROJECT2 ;

VAR Vehicles\_involved;

WITH Victims;

TITLE "correlation between vehicles involved and victims";

**RUN**;

ODS RTF CLOSE;

\*both serious and mild injuries are correlated with the number of victims .71 nd .97;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **CORR** DATA=CAR.PROJECT2 ;

VAR Serious\_injuries Mild\_injuries;

WITH Victims;

TITLE "correlation between serious and mild injuries with victims";

**RUN**;

ODS RTF CLOSE;

\*CATEGORICAL COLUMNS;

\*CHISQ;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **FREQ** DATA = CAR.PROJECT2;

TABLE District\_Name\* Weekday/CHISQ NOCOL NOROW ;

TITLE "CHISQ ANALYSIS FOR DISTRICT NAME BY WEEKDAY";

**RUN**;

ODS RTF CLOSE;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **FREQ** DATA = CAR.PROJECT2;

TABLE Month\* Part\_of\_the\_day/CHISQ NOCOL NOROW ;

TITLE "CHISQ ANALYSIS FOR MONTH AND PART OF THE DAY";

**RUN**;

ODS RTF CLOSE;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **FREQ** DATA = CAR.PROJECT2;

TABLE Weekday\*Part\_of\_the\_day/CHISQ NOCOL NOROW ;

TITLE "CHISQ ANALYSIS FOR PART OF THE DAY AND WEEKDAY";

**RUN**;

ODS RTF CLOSE;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**Proc** **freq** data=CAR.PROJECT2;

Tables Month Hour Weekday Part\_of\_the\_day;

title"MONTH,TIME AND PART OF THE DAY FREQUENCIES";

**Run**;

ods rtf close;

\*PRODUCING STATISTICS WITH PROC MEANS;

\*The average number of victims was 1.18.The median value of 1 says that the victims sampled had 1 death;

\*the confidence limit tell us that we are 95% certain that the true population mean falls between 1.16 & 1.19 ;

\*conclusion is that most accidents victims on average are approximately 2;

**PROC** **MEANS** DATA = CAR.PROJECT2 MAXDEC= **2** N NMISS MIN MEAN Median STD MAX CLM STDERR;

VAR Latitude Longitude Mild\_injuries Serious\_injuries Vehicles\_involved Victims;

**RUN**;

\*OUTLIER DETECTION;

\*PRODUCING STATISTICS WITH PROC MEANS;

**PROC** **MEANS** DATA = CAR.PROJECT2 MAXDEC= **2** N NMISS MIN MEAN Median q1 q3 STD MAX ;

VAR Latitude Longitude Mild\_injuries Serious\_injuries Vehicles\_involved Victims;

**RUN**;

\*PROC ANOVA;

**PROC** **ANOVA** DATA = CAR.PROJECT2;

CLASS Hour;

MODEL Longitude = Hour;

MEANS Hour/Scheffe;

TITLE "Longitude by hour";

**RUN**;

**PROC** **ANOVA** DATA = CAR.PROJECT2;

CLASS Hour;

MODEL Longitude = Hour;

MEANS Hour/Scheffe;

TITLE "Longitude by hour";

**RUN**;

**PROC** **ANOVA** DATA = CAR.PROJECT2;

CLASS Hour;

MODEL Latitude = Hour;

MEANS Hour/Scheffe;

TITLE "Longitude by hour";

**RUN**;

\*because the p value is 0.0237 we conclude that there are no difference in means of victims and the three parts of the day;

**PROC** **ANOVA** DATA = CAR.PROJECT2;

CLASS Part\_of\_the\_day;

MODEL Victims = Part\_of\_the\_day;

MEANS Part\_of\_the\_day/Scheffe;

TITLE "Victims by Part\_of\_the\_day";

**RUN**;

\*because the p value is 0.6471 we conclude that there are is difference in means of vehicles involved and the three parts of the day;

**PROC** **ANOVA** DATA = CAR.PROJECT2;

CLASS Part\_of\_the\_day;

MODEL Vehicles\_involved = Part\_of\_the\_day;

MEANS Part\_of\_the\_day/Scheffe;

TITLE "Victims by Part\_of\_the\_day";

**RUN**;

\*proc ttest;

\*the mean differene betwn serious injuries and victims is -1.1553.the ttest shows significant evidence (<.0001)of a diff betwween the mean of serious and victims;

\*conclusion no significance between the serius injured and victims;

**proc** **ttest** data=CAR.PROJECT2;

TITLE"MEAN DIFF BTW SERIOUS INJURED AND VICTIMS";

PAIRED Serious\_injuries \*Victims;

**run**;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**proc** **ttest** data=CAR.PROJECT2;

var Victims;

**run**;

ODS RTF CLOSE;

\*proc univariate;

\*the relatively linear pattern formed by the points in the probability plot indicate that the data are closely matched to normal distribution;

**PROC** **UNIVARIATE** DATA = CAR.PROJECT2;

VAR Vehicles\_involved;

HISTOGRAM Vehicles\_involved /NORMAL;

PROBPLOT Vehicles\_involved;

TITLE"PROB PLOT FOR VICTIMS";

**RUN**;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**proc** **univariate** data=CAR.PROJECT2 normal plot freq;

var Victims;

output out=sumdata

n=n min=min

mean=mean std=std

median=median;

TITLE"VICTIMS PER CLASH";

**run**;

ODS RTF CLOSE;

**PROC** **SQL** OUTOBS=**10**;

SELECT District\_Name,Street,SUM(Victims)AS Loss\_LIFE

FROM CAR.PROJECT2

GROUP BY District\_Name,Street

ORDER BY Loss\_LIFE DESC

;

**QUIT**;

TITLE"MOST DANGEROUS STREETS";

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**PROC** **SQL** OUTOBS=**10**;

SELECT District\_Name,Neighborhood\_Name, SUM(Victims)AS DEATHS

FROM CAR.PROJECT2

GROUP BY District\_Name,Neighborhood\_Name

ORDER BY DEATHS DESC

;

**QUIT**;

ODS RTF CLOSE;

**PROC** **SQL** OUTOBS=**5**;

SELECT District\_Name, SUM(Victims)AS victims\_desc

FROM CAR.PROJECT2

GROUP BY District\_Name

ORDER BY victims\_desc DESC

;

**QUIT**;

**PROC** **SQL**;

SELECT District\_Name,Neighborhood\_Name, SUM(Victims)AS DEATHS

FROM CAR.PROJECT2

GROUP BY District\_Name,Neighborhood\_Name

ORDER BY DEATHS DESC

;

**QUIT**;

ODS RTF CLOSE;

**proc** **format**;

VALUE VGRP **0**-**2**="<2"

**3**-**4**="3-4"

**5**-**6**="5-6"

**7**-**8**="7-8"

**9**-**10**=">8";

TITLE"Victims Distribution";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Victims;

FORMAT Victims VGRP.;

**RUN**;

**proc** **format**;

VALUE hourgrp **0**-**6**="0-6"

**7**-**12**="7-12"

**13**-**18**="13-18"

**19**-**24**="19-24";

TITLE"ACCIDENTS PER HOUR GROUP";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Hour;

FORMAT Hour hourgrp.;

**RUN**;

**proc** **format**;

VALUE Vinvolved **0**-**2**="<2"

**3**-**4**="3-4"

**5**-**6**="5-6"

**7**-**8**="7-8"

**9**-**10**="9-10"

**11**-**12**="11-12"

**13**-**14**="13-14"

TITLE"VEHICLES INVOLVED GROUPED";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Vehicles\_involved;

FORMAT Vehicles\_involved Vinvolved.;

**RUN**;

**proc** **format**;

VALUE VGRP **0**-**2**="<2"

**3**-**4**="3-4"

**5**-**6**="5-6"

**7**-**8**="7-8"

**9**-**10**=">8";

TITLE"Distribution of Victims & Part of the Day";

**RUN**;

**PROC** **SGPLOT** DATA=CAR.PROJECT2;

VBAR Victims/GROUP=Part\_of\_the\_day;

FORMAT Victims VGRP.;

**RUN**;

**data** CAR.PROJECT2;

Part\_of\_the\_day="Night";

attrib Part\_of\_the\_day format=$10.;

**run**;

**PROC** **SUMMARY** data=CAR.PROJECT2;

VAR Victims Mild\_injuries Serious\_injuries Vehicles\_involved;

output out=sumdata;

**run**;

**proc** **print** data=sumdata;

**run**;

**PROC** **TABULATE** DATA= CAR.PROJECT2;

CLASS District\_Name Neighborhood\_Name ;

VAR Victims;

TABLE District\_Name\*Neighborhood\_Name /

RTS=**5**;

**RUN**;

**PROC** **TABULATE** DATA= CAR.PROJECT2;

CLASS Neighborhood\_Name ;

VAR Victims;

TABLE Neighborhood\_Name\*Victims /

RTS=**20**;

**RUN**;

ODS RTF FILE="C:\Users\User\Desktop\CAR.RTF";

**proc** **reg** data=CAR.PROJECT2;

TITLE"REGRESSION ANALYSIS ";

model Victims =Vehicles\_involved;

title"REGRESSION ANALYSIS MODEL"

run;

ods rtf close;

**proc** **sgmap** plotdata=CAR.PROJECT;

openstreetmap;

scatter x=Longitude y=Latitude /group=Part\_of\_the\_day markerattrs=(symbol=circlefilled size=**8**);

title"Barcelona Accidents Locations and Part of the Day";

**run**;

**proc** **sgmap** plotdata=CAR.PROJECT;

openstreetmap;

scatter x=Longitude y=Latitude /group=Victims markerattrs=(symbol=circlefilled size=**8**);

title"Barcelona Accidents Locations Sgmap";

**run**;