Seattle Occupies List of World's Worst Traffic Cities

(Here a rollover crash blocks all southbound lanes of Interstate 5, causing long traffic backups)

Predicting Traffic Collision Severity i
in Seattle, Washington, USA

WHAT ARE THE CHANCES YOU WILL BE INJURED OR KILLED IN A SEATTLE TRAFFIC ACCIDENT?

Author

Jerry W. Hedgepath

TABLE OF CONTENTS

INTRODUCTION/BUSINESS PROBLEM STATEMENT

SEATTLE TRAFFIC BACKGROUND

According to Joel Connelly, a reporter for the *Seattle Post-Intelligencer*, Seattle ranks as number 20 in a recent CBS News compilation of “Cities with the worst traffic in the world.” Seattle is just ahead of Dallas and St Petersburg in Russia, and trails just behind Chicago and Boston. The striking feature of the list is that almost all of the cities on the list are larger than Seattle. The top five cities are: 1) Los Angeles; 2) Moscow; 2 tie) New York City; 4) Sao Paulo, Brazil; and 5) San Francisco, CA.

Seattle’s ranking is a product of three “G’s – Geography, Growth and Guilt.

Seattle, long known as the Emerald City for lush forests surrounding the city, is squeezed between two bodies of water, Elliott Bay and Lake Washington. It has just two major north-south highways, Interstate 5 and State Route 99.

It also features world-class examples of engineering ineptitude, such as drivers coming off state Route 520 (the Evergreen Point Bridge), joining southbound I-5 in the left lane, and having less than a mile to cross four lanes of freeway to exit on Mercer. And vice versa.

Seattle has gained more than 100,000 new residents in the past eight years. Cities north, south and east are growing as well.

The guilt? Seattle-area voters twice turned down, in the late 1960's, a proposed rail system. Sen. Warren Magnuson had secured federal money to pay the bulk of the bill. Sadly, the city's construction unions were addicted to concrete, and led the opposition.

On average, Seattle drivers each lost 55 hours to traffic during peak times.

PROBLEM STATEMENT

With the traffic problems outlined above, the ability to accurately analyze and model traffic accident data becomes increasingly important. A baseline ability to predict the “seriousness” of a future accident is key along with drawing insights into traffic patterns based on time of day, day of week, weather, lighting and road conditions, and other attributes. Additionally, a variety of insights may be derived to benefit urban planning efforts and improving transportation infrastructure.

DATA SOURCES

SEATTLE DEPARTMENT OF TRANSPORTATION TRAFFIC DATA

The homepage of the Seattle Department of Transportation traffic data is:

http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv

TRAFFIC DATA METADATA

Meta-data of the dataset can be viewed at
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

TRAFFIC DATA ANALYSIS

The labelled dataset contains 221,389 data rows. The dataset was last updated on September 5, 2020 and accessed on September 18, 2020. The dataset covers the time frame from 2004 to last update date. The dataset contains 40 attributes some of which may not be useful for modeling.

Because the dataset it updated frequently and to provide a stable basis for analysis, a copy of the dataset was uploaded to IBM Cloud Storage. This copy of the dataset was used in the following analysis.

Some basic information from the dataframe:

a. Head (partial column display)

```
In [4]: #read the data
        pd.set_option('display.max_columns', None)
        df = pd.read_csv("http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv")
        print("Data read into dataframe!")

        Data read into dataframe!

In [5]: # A cursory analysis of the data
        df.head()
```

Out[5]:

	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDESC	SEVERITYCODE	SEVERITYDESC	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
0	-122.320757	47.609408	1	328476	329976	EA08706	Matched	Block	NaN	BROADWAY BETWEEN E COLUMBIA ST AND BOYLSTON AVE		NaN	1	Property Damage Only Collision	Sideswipe	2	0	0	2
1	-122.319561	47.662221	2	328142	329642	EA06882	Matched	Block	NaN	8TH AVE NE BETWEEN NE 45TH E ST AND NE 47TH ST		NaN	1	Property Damage Only Collision	Parked Car	2	0	0	2
2	-122.327525	47.604393	3	20700	20700	1181833	Unmatched	Block	NaN	JAMES ST BETWEEN 6TH AVE AND 7TH AVE	NaN	NaN	0	Unknown	NaN	0	0	0	0
3	-122.327525	47.708622	4	332126	333626	M16001640	Unmatched	Block	NaN	NE NORTHGATE WAY BETWEEN 1ST AVE NE AND NE NOR...		NaN	0	Unknown	NaN	0	0	0	0
4	-122.292120	47.559009	5	328238	329738	3857118	Unmatched	Block	NaN	M L KING JR ER WAY S BETWEEN S ANGELINE ST AND...		NaN	0	Unknown	NaN	0	0	0	0

b. Shape and column values

```
In [7]: df.shape

Out[7]: (221389, 40)

In [9]: df.columns.values

Out[9]: array(['X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS',
               'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC',
               'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT',
               'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES',
               'SERIOUSINJURIES', 'FATALITIES', 'INCDATE', 'INCDTTM',
               'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'INATTENTIONIND',
               'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT',
               'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY',
               'CROSSWALKKEY', 'HITPARKEDCAR'], dtype=object)
```

c. Datatypes

```
In [12]: df.dtypes

Out[12]: X                   float64
         Y                   float64
         OBJECTID              int64
         INCKEY                int64
         COLDETKEY             int64
         REPORTNO             object
         STATUS               object
         ADDRTYPE             object
         INTKEY              float64
         LOCATION             object
         EXCEPTRSNCODE        object
         EXCEPTRSNDESC        object
         SEVERITYCODE         object
         SEVERITYDESC         object
         COLLISIONTYPE        object
         PERSONCOUNT           int64
         PEDCOUNT              int64
         PEDCYLCOUNT           int64
         VEHCOUNT              int64
         INJURIES              int64
         SERIOUSINJURIES       int64
         FATALITIES            int64
         INCDATE              object
         INCDTTM              object
         JUNCTIONTYPE         object
         SDOT_COLCODE        float64
         SDOT_COLDESC         object
         INATTENTIONIND       object
         UNDERINFL            object
         WEATHER              object
         ROADCOND             object
         LIGHTCOND            object
         PEDROWNOTGRNT        object
         SDOTCOLNUM          float64
         SPEEDING             object
         ST_COLCODE           object
         ST_COLDESC           object
         SEGLANEKEY            int64
         CROSSWALKKEY          int64
         HITPARKEDCAR         object
         dtype: object
```

d. Datatypes counts

```
In [11]: df.dtypes.value_counts()

Out[11]: object     23
         int64      12
         float64     5
         dtype: int64
```

METHODOLOGY SECTION

DATA ANALYSIS & CLEANSING

At the start of this process, the dataset contain 221,389 rows with 40 attributes. There are a number of attributes which are not useful for further analysis. The following table lists the attributes which were dropped.

	Dropped Attributes
Attribute	**Description**
OBJECTID	ESRI Unique Identifier
INCKEY	Unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	Description not available
STATUS	Description not available
INTKEY	Collision intersection key
SDOT_COLCODE	SDOT collision code
SDOT_COLDESC	SDOT collision description
ST_COLCODE	Washington State collision code
ST_COLDESC	Washington State collision description
SEGLANEKEY	Lane segment key
CROSSWALKKEY	Crosswalk key
SDOTCOLNUM	SDOT collision number

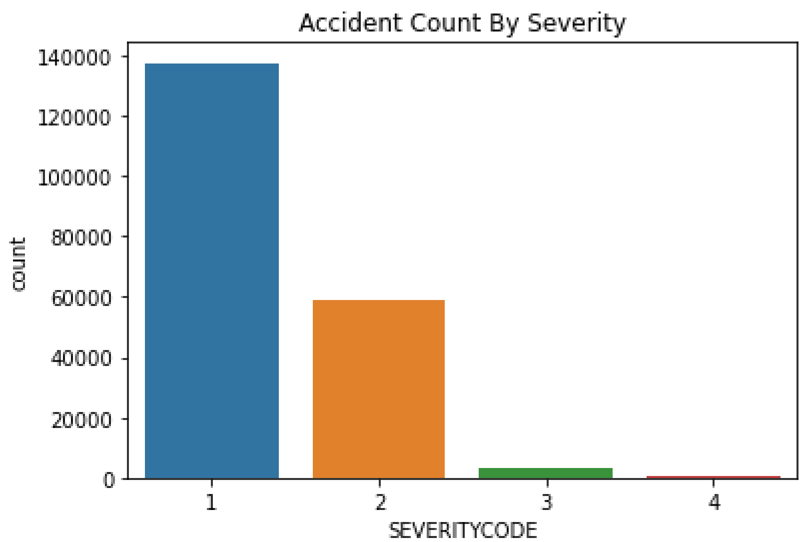
Our target attribute, SEVERITYCODE, has several values. One of which is “0 – Unknown”. Since “Unknown” severity codes have no value in our analysis, 21,595 rows were identified and deleted. After these deletions, 199,794 rows with 27 columns remained for further processing.

SEVERITYCODE now contains the following values: 1-property damage only, 2-injury, 2b-serious injury, 3-fatality. In order to allow for further analysis, SEVERITYCODE was realigned so that 3 --> 4, and 2b --> 3.

Once this realignment is complete, SEVERITYCODE breakdown is:

Code	Description	Percent
1	Property Damage Only	68.9
2	Injury	29.4
3	Serious Injury	1.6
4	Fatality	0.2

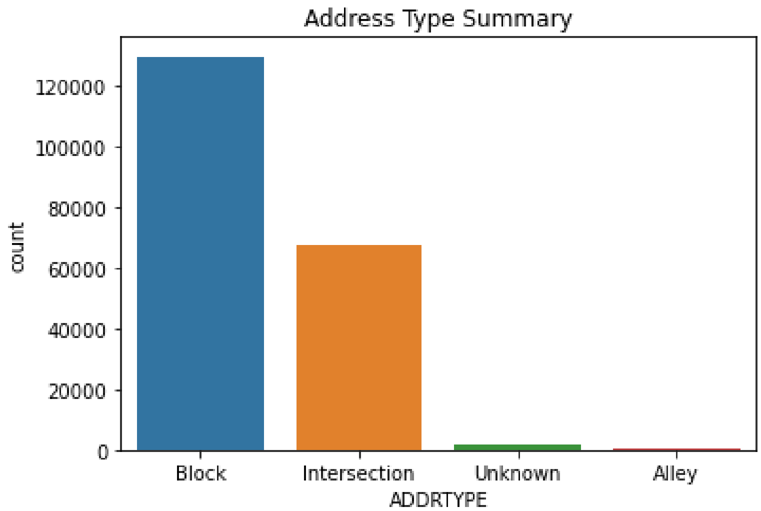
Almost 69% of collisions involved property damage only, no injuries or fatalities.



The ADDRTYPE column contains many blank rows. They were cleaned and set to "Unknown". The ADDRTYPE breakdown is:

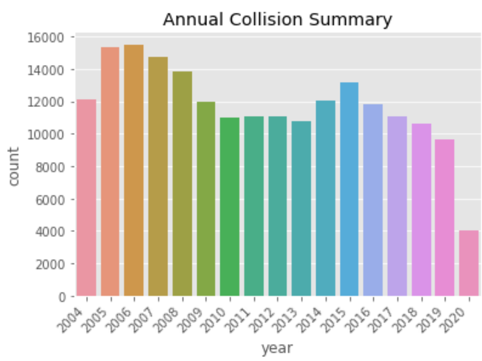
Address Type	Percent
Block	64.9
Intersection	33.7
Unknown	1.0
Alley	0.4

Almost 65% of collisions occurred within city block limits.

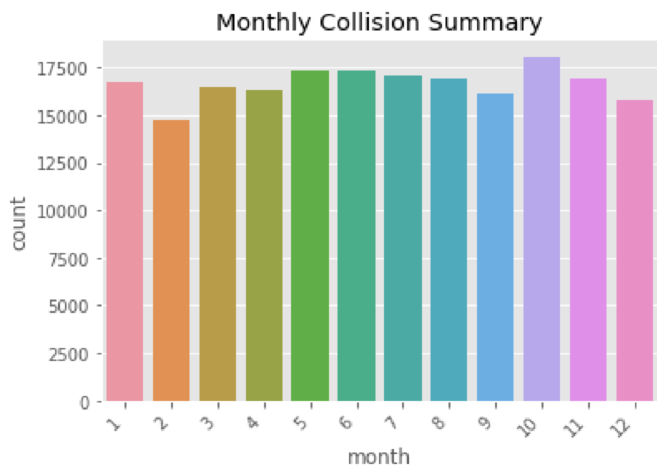


INCDATE and INCDTTM columns were changed to datetime format. Which led to the ability to extract year, month, and weekday from INCDTTM and add them as columns to the dataframe.

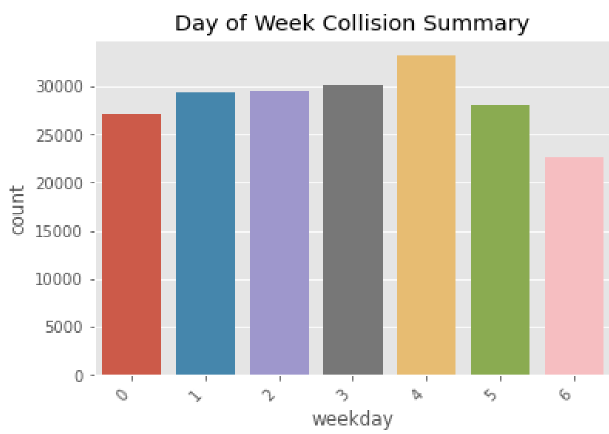
Let's look at the distribution over years, months, and weekdays.



2020 shows a large decrease in collisions. It is reasonable to assume this is due to the Covid-19 pandemic which had Seattle drivers sheltering in place for several weeks, thereby reducing traffic volumes and collisions. Additionally, 2020 contains only 9 1/2 months of data, since data was pulled on September 17, 2020.



Graph indicated higher collision count during the winter months of October through January.

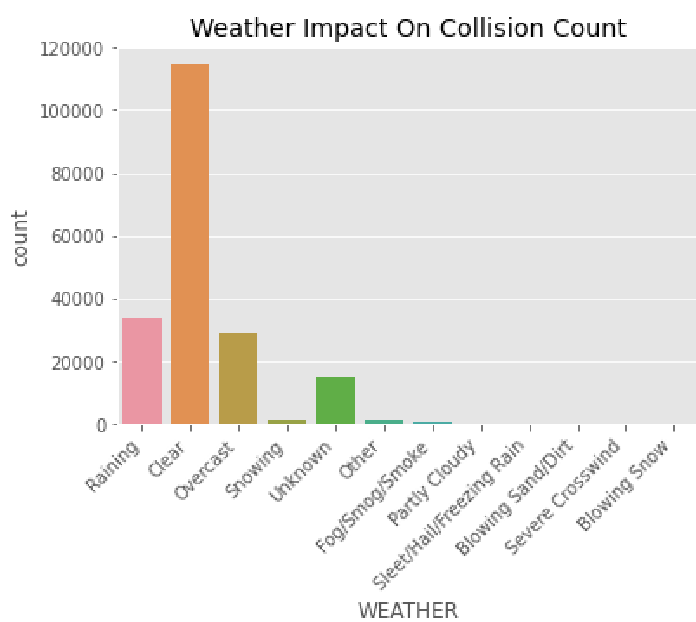


0 = Monday; 1 = Tuesday; 2 = Wednesday; 3 = Thursday; 4 = Friday; 5 = Saturday; 6 = Sunday
Graph indicates increased traffic collisions during Thursday-Saturday period.

How do weather conditions impact collisions.

Weather Condition	Percent
Clear	58.8
Raining	17.5
Overcast	14.6
Unknown	7.8
Snowing	0.5
Other	0.4
Sleet/Hail/Freezing Rain	0.1
Blowing Sand/Dirt	0.0
Severe Crosswind	0.0
Partly Cloudy	0.0
Blowing Snow	0.0

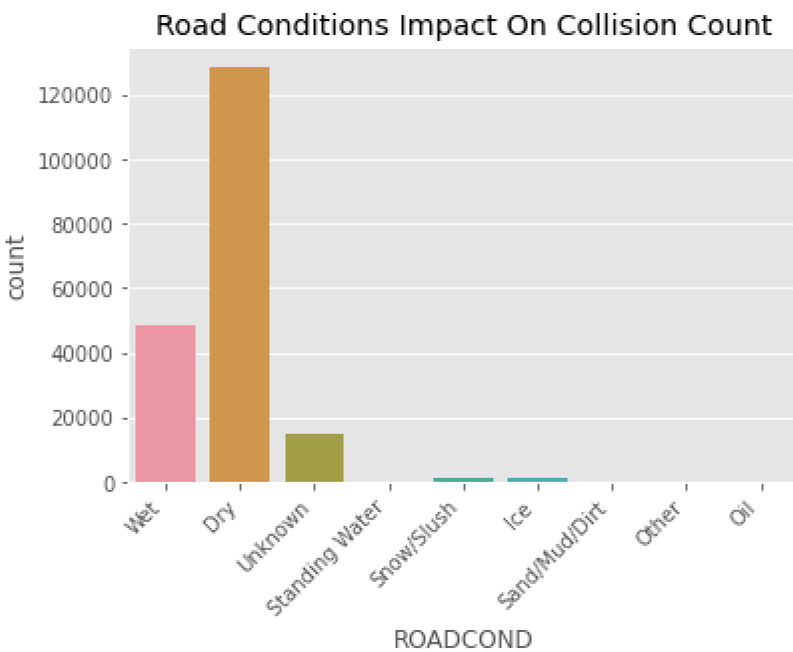
Almost 59% of collisions occurred during clear weather.



How did road conditions impact collisions?

Road Condition	Percent
Dry	65.9
Wet	25.0
Unknown	7.8
Ice	0.6
Snow/Slush	0.5
Other	0.1
Standing Water	0.1
Sand/Mud/Dirt	0.0
Oil	0.0

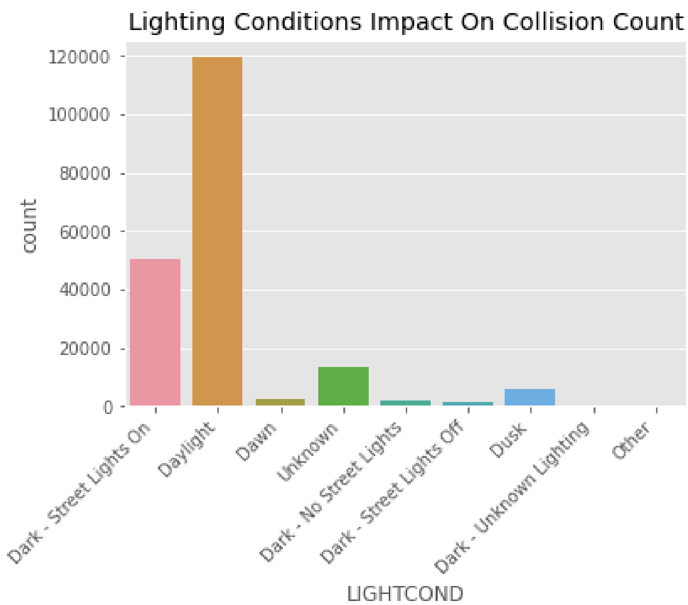
Over 65% of collisions occurred during dry road conditions.



What was the impact of light conditions on collisions?

Light Conditions	Percent
Daylight	61.3
Dark-Street Lights On	25.7
Unknown	6.9
Dusk	3.1
Dawn	1.3
Dark-No Street Lights	0.8
Dark-Street Lights Off	0.6
Other	0.1
Dark-Unknown Lighting	0.0

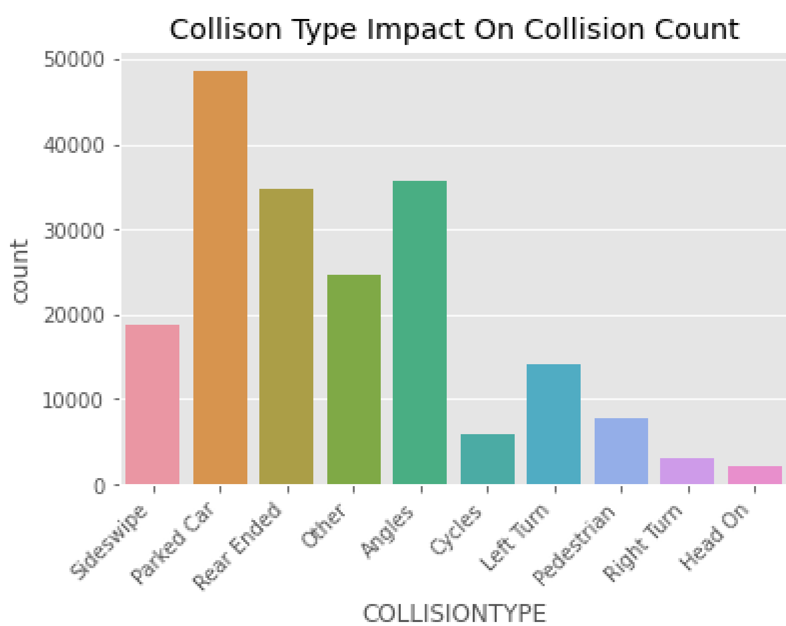
Over 61% of collisions occurred during daylight.



Let's look at Collision Type.

Collision Type	**Percent**
Parked Car	24.9
Angles	18.2
Rear Ended	17.8
Other	12.6
Sideswipe	9.7
Left Turn	7.2
Pedestrian	3.9
Cycles	3.0
Right Turn	1.5
Head On	1.1

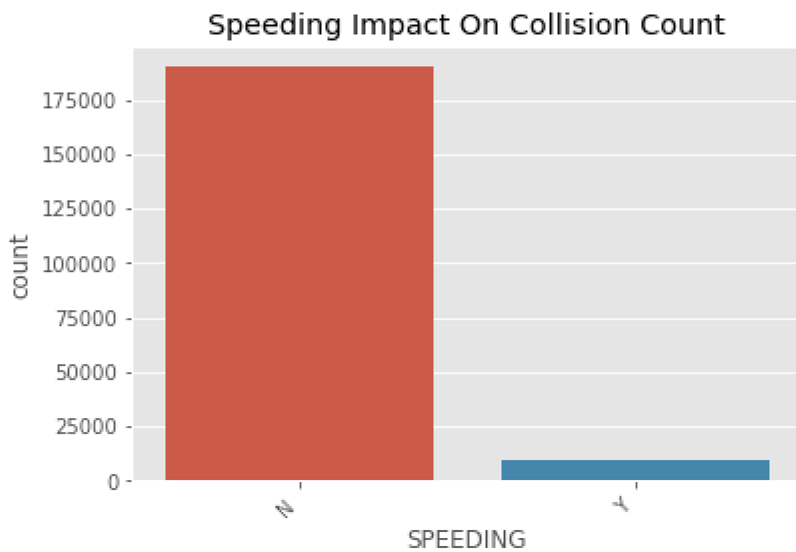
Almost 25% of collisions involved hitting a parked car.



How often was speeding a contributing factor to a collision?

Speeding	**Percent**
No	95.03
Yes	4.97

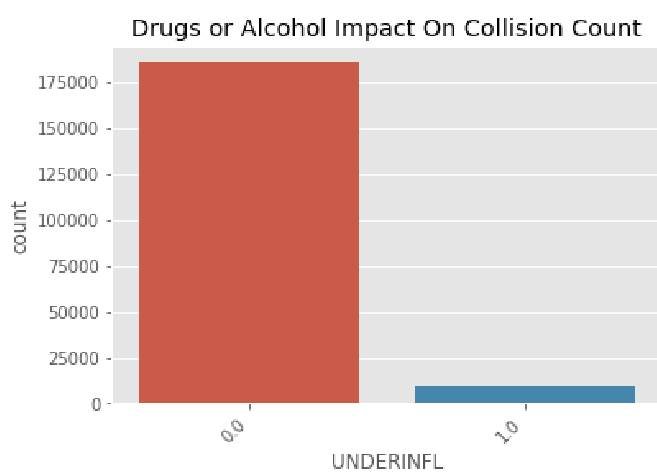
Less than 5% of collisions had speeding as a contributing factor.



Let's look at Under the Influence (UNDERINFL). After cleaning up the data and setting UNDERINFL = 1, driver was under the influence; and 0 = driver was not under the influence of drugs or alcohol.

Under the Influence	**Percent**
0=Not under influence	95.07
1=Under the influence	4.93

Over 95% of collisions did not involve a driver under the influence of drugs or alcohol.



In order to utilize the LabelEncoder module, certain attributes need to be cleaned (i.e. handle missing values, handle NaN values). These attributes include WEATHER, ROADCOND, LIGHTCOND, COLLISIONTYPE, and UNDERINFL.

The following table lists the original attributes processed with the LabelEncoder module and the related new columns added to the dataframe.

LabelEncoder		
Attribute	**Description**	**LabelEncoder Name**
ADDRTYPE	Collision address type	le_addr
WEATHER	Weather conditions	le_weather
ROADCOND	Road conditions	le_roadcond
LIGHTCOND	Lighting conditions	le_lightcond
COLLISIONTYPE	Collision type	le_collisiontype
SPEEDING	Speeding a collision factor	le_speeding
UNDERINFL	Driver under the influence of drugs or alcohol	le_underinfl

The dataset is severely unbalanced with SEVERITYCODE=1 (Property damage only) greatly outnumbering the other SEVERITYCODE counts, as indicated here:

```
1    137596
2     58747
3      3102
4       349
Name: SEVERITYCODE, dtype: int64
```

We can correct the imbalance by using the SMOTE module in the imblearn library. SMOTE(Synthetic Minority Oversampling Technique) uses the K-Nearest Neighbors algorithm to synthetically generated examples for the minority classes (SEVERITYCODE 2-3-4). We sample up to generate minority classes examples in line with the majority class, as indicated here:

```
4.0    137596
3.0    137596
2.0    137596
1.0    137596
Name: SEVERITYCODE, dtype: int64
```

At this point, we are ready to split the collision data, utilizing the datetime fields and LabelEncoder fields added to the dataframe.

Let's split the collision data; SEVERITYCODE is our target variable.

X = df[['le_addr', 'year', 'month', 'weekday', 'le_weather', 'le_roadcond', 'le_lightcond', 'le_collisiontype', 'le_speeding', 'le_underinfl']]

y = df['SEVERITYCODE']

X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.2, random_state=1)

print ('Train set:', X_train.shape, y_train.shape)

print ('Test set:', X_test.shape, y_test.shape)

```
Train set: (440307, 10) (440307,)
Test set: (110077, 10) (110077,)
```

After SMOTE. SEVERITYCODE values are more evenly distributed across the four values:

```
4    110286
1    110048
2    110042
3    109931
Name: SEVERITYCODE, dtype: int64
```

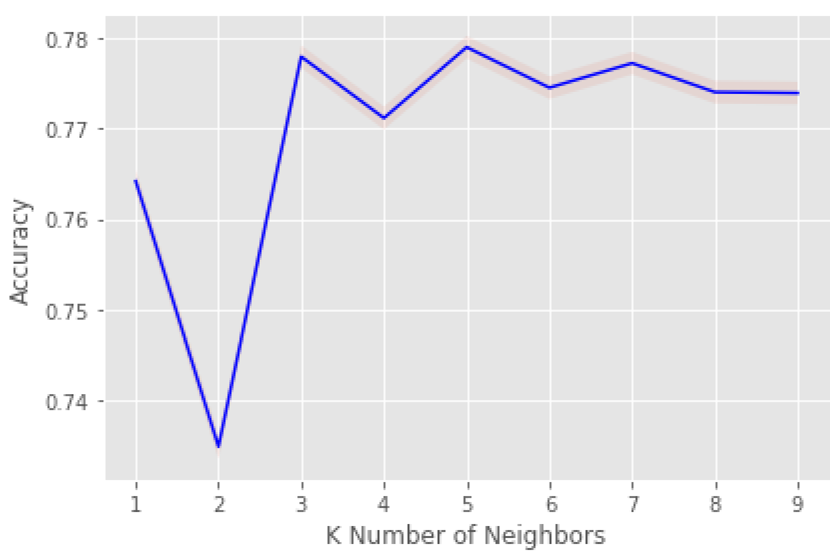
RESULTS SECTION

K-NEAREST NEIGHBORS

Let's start the analysis with K-Nearest Neighbors. With Ks=10, the following results array is generated:

```
array([0.7642, 0.7348, 0.778 , 0.7712, 0.779 , 0.7745, 0.7773, 0.7741,
       0.774 ])
```

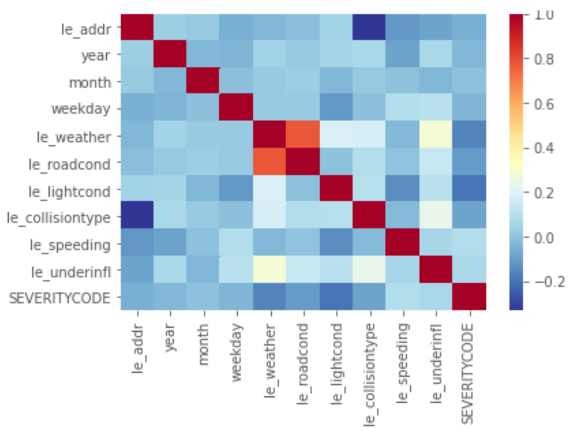
When mapped:



In this example, K=5, with an accuracy of 77.9 %.

HEAT MAP

Let's examine a heatmap which illustrates the correlation between certain variables.



In the map above, red indicate a positive or strong correlation between the two variables; while blue indicates a negative or weak correlation.

The map shows little positive correlation and a large amount of negative correlation.

LOGISTIC REGRESSION ANALYSIS

Here's the Classification Report:

	precision	recall	f1-score	support
1	0.37	0.47	0.42	27548
2	0.30	0.21	0.25	27554
3	0.32	0.30	0.31	27665
4	0.38	0.44	0.41	27310
accuracy			0.35	110077
macro avg	0.35	0.35	0.34	110077
weighted avg	0.35	0.35	0.34	110077

DECISION TREE ANALYSIS

Here's the Classification Report:

	precision	recall	f1-score	support
1	0.60	0.49	0.54	27548
2	0.44	0.58	0.50	27554
3	0.47	0.08	0.14	27665
4	0.41	0.72	0.52	27310
accuracy			0.46	110077
macro avg	0.48	0.46	0.43	110077
weighted avg	0.48	0.46	0.42	110077

RANDOM FOREST CLASSIFIER

Here's the Classification Report:

	precision	recall	f1-score	support
1	0.76	0.64	0.70	27548
2	0.69	0.70	0.70	27554
3	0.85	0.93	0.89	27665
4	0.94	0.99	0.96	27310
accuracy			0.81	110077
macro avg	0.81	0.81	0.81	110077
weighted avg	0.81	0.81	0.81	110077

DISCUSSION SECTION

The three Classification reports are summarized in the following table:

SEVERITYCODE	Logistic Regression			Decision Tree			Random Forest		
	Precision	Recall	F1-score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1-Property Damage Only	0.37	0.47	0.42	0.60	0.49	0.54	0.76	0.64	0.70
2-Injury	0.30	0.21	0.25	0.44	0.58	0.50	0.69	0.70	0.70
3-Serious Injury	0.32	0.30	0.31	0.47	0.08	0.14	0.85	0.93	0.89
4-Fatality	0.38	0.44	0.41	0.41	0.72	0.52	0.94	0.99	0.96
accuracy			0.35			0.46			0.81
macro avg	0.35	0.35	0.34	0.48	0.46	0.43	0.81	0.81	0.81
weighted avg	0.35	0.35	0.34	0.48	0.46	0.42	0.81	0.81	0.81

Given the data utilized in training and testing the various models, the Random Forest Classifier provides the most accurate predictions, especially collisions involving Serious Injury and Fatalities. On average, The Random Forest Classifier is 2.3 times more accurate than Logistic Regression and 1.7 times more accurate than Decision Tree Analysis in predicting the seriousness of a collision.

CONCLUSION SECTION

The data attributes, tools, and techniques used in this report produced valuable insights into predicting the seriousness of a collision. That is not to say, that utilizing other data attributes, tools, and techniques would produce better results.