This notebook contains an expainaton of the data that we will be using for our Capstone Project.

Importing and Libraries Dataset

In [1]:

```
# Importing Only Essential Libraries
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
pd.options.plotting.backend = "plotly"
```

In [2]:

```
# Importing Data
dataset = pd.read_csv('Data_Collisions.csv', low_memory = False)
dataset.head(10)
```

Out[2]:

	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTY
0	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersect
1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Bl
2	-122.334540	47.607871	3	26700	26700	1482393	Matched	Bl
3	-122.334803	47.604803	4	1144	1144	3503937	Matched	Bl
4	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersect
5	-122.387598	47.690575	6	320840	322340	E919477	Matched	Intersect
6	-122.338485	47.618534	7	83300	83300	3282542	Matched	Intersect
7	-122.320780	47.614076	9	330897	332397	EA30304	Matched	Intersect
8	-122.335930	47.611904	10	63400	63400	2071243	Matched	Bl
9	-122.384700	47.528475	12	58600	58600	2072105	Matched	Intersect

10 rows × 37 columns

Data Preprocessing

In [3]:

```
print('Number of Rows:',dataset.shape[0])
print('Number of Columns:',dataset.shape[1])
```

```
Number of Rows: 194673
Number of Columns: 37
```

As seen above this dataset contains 194673 Rows and 38 Columns. But, it also contains some Nan values which we need to remove from the data in order to get a perfect model accuracy.

Notice that in the data I have moved the Severity column to the very last just for the ease of analysing; as you will see below.

In [4]:

```
dataset.isnull().sum()
```

Out[4]:

```
X                   5334
Y                   5334
OBJECTID               0
INCKEY                 0
COLDETKEY              0
REPORTNO               0
STATUS                 0
ADDRTYPE            1926
INTKEY            129603
LOCATION            2677
EXCEPTRSNCODE     109862
EXCEPTRSNDESC     189035
SEVERITYDESC           0
COLLISIONTYPE       4904
PERSONCOUNT            0
PEDCOUNT               0
PEDCYLCOUNT            0
VEHCOUNT               0
INCDATE                0
INCDTTM                0
JUNCTIONTYPE        6329
SDOT_COLCODE           0
SDOT_COLDESC           0
INATTENTIONIND    164868
UNDERINFL           4884
WEATHER             5081
ROADCOND            5012
LIGHTCOND           5170
PEDROWNOTGRNT     190006
SDOTCOLNUM         79737
SPEEDING          185340
ST_COLCODE            18
ST_COLDESC          4904
SEGLANEKEY             0
CROSSWALKKEY           0
HITPARKEDCAR           0
SEVERITY               0
dtype: int64
```

In [5]:

```
# Dropping all the irrelevant rows
dataset.drop(['OBJECTID', 'REPORTNO', 'STATUS','PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'INATTENTIONIND', 'INTKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC'], axis = 1, inplace = True)
```

In [6]:

```
dataset.isnull().sum()
```

Out[6]:

```
X                  5334
Y                  5334
INCKEY                0
COLDETKEY             0
ADDRTYPE           1926
LOCATION           2677
SEVERITYDESC          0
COLLISIONTYPE      4904
PERSONCOUNT           0
PEDCOUNT              0
PEDCYLCOUNT           0
VEHCOUNT              0
INCDATE               0
INCDTTM               0
JUNCTIONTYPE       6329
SDOT_COLCODE          0
SDOT_COLDESC          0
UNDERINFL          4884
WEATHER            5081
ROADCOND           5012
LIGHTCOND          5170
ST_COLCODE           18
ST_COLDESC         4904
SEGLANEKEY            0
CROSSWALKKEY          0
HITPARKEDCAR          0
SEVERITY              0
dtype: int64
```

In [7]:

```
# Dropping all NaN Values
dataset.dropna(axis = 0, inplace = True)
```

In [8]:

```
dataset.isnull().sum()
```

Out[8]:

```
X                0
Y                0
INCKEY           0
COLDETKEY        0
ADDRTYPE         0
LOCATION         0
SEVERITYDESC     0
COLLISIONTYPE    0
PERSONCOUNT      0
PEDCOUNT         0
PEDCYLCOUNT      0
VEHCOUNT         0
INCDATE          0
INCDTTM          0
JUNCTIONTYPE     0
SDOT_COLCODE     0
SDOT_COLDESC     0
UNDERINFL        0
WEATHER          0
ROADCOND         0
LIGHTCOND        0
ST_COLCODE       0
ST_COLDESC       0
SEGLANEKEY       0
CROSSWALKKEY     0
HITPARKEDCAR     0
SEVERITY         0
dtype: int64
```

In [9]:

```
dataset = pd.DataFrame(dataset)
dataset.head(10)
```

Out[9]:

	X	Y	INCKEY	COLDETKEY	ADDRTYPE	LOCATION	SEVERITYDESC
0	-122.323148	47.703140	1307	1307	Intersection	5TH AVE NE AND NE 103RD ST	Injury Collision
1	-122.347294	47.647172	52200	52200	Block	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	Property Damage Only Collision
2	-122.334540	47.607871	26700	26700	Block	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	Property Damage Only Collision
3	-122.334803	47.604803	1144	1144	Block	2ND AVE BETWEEN MARION ST AND MADISON ST	Property Damage Only Collision
4	-122.306426	47.545739	17700	17700	Intersection	SWIFT AVE S AND SWIFT AV OFF RP	Injury Collision
5	-122.387598	47.690575	320840	322340	Intersection	24TH AVE NW AND NW 85TH ST	Property Damage Only Collision
6	-122.338485	47.618534	83300	83300	Intersection	DENNY WAY AND WESTLAKE AVE	Property Damage Only Collision
7	-122.320780	47.614076	330897	332397	Intersection	BROADWAY AND E PIKE ST	Injury Collision
8	-122.335930	47.611904	63400	63400	Block	PINE ST BETWEEN 5TH AVE AND 6TH AVE	Property Damage Only Collision
9	-122.384700	47.528475	58600	58600	Intersection	41ST AVE SW AND SW THISTLE ST	Injury Collision

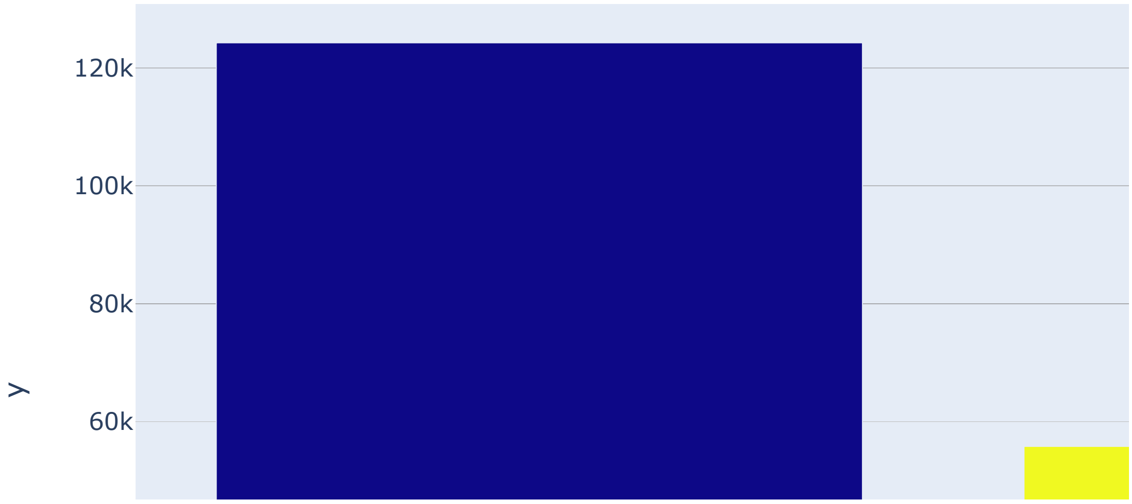
10 rows × 27 columns

Data Visualization

Now that our data is cleaned properly, lets visualize it.

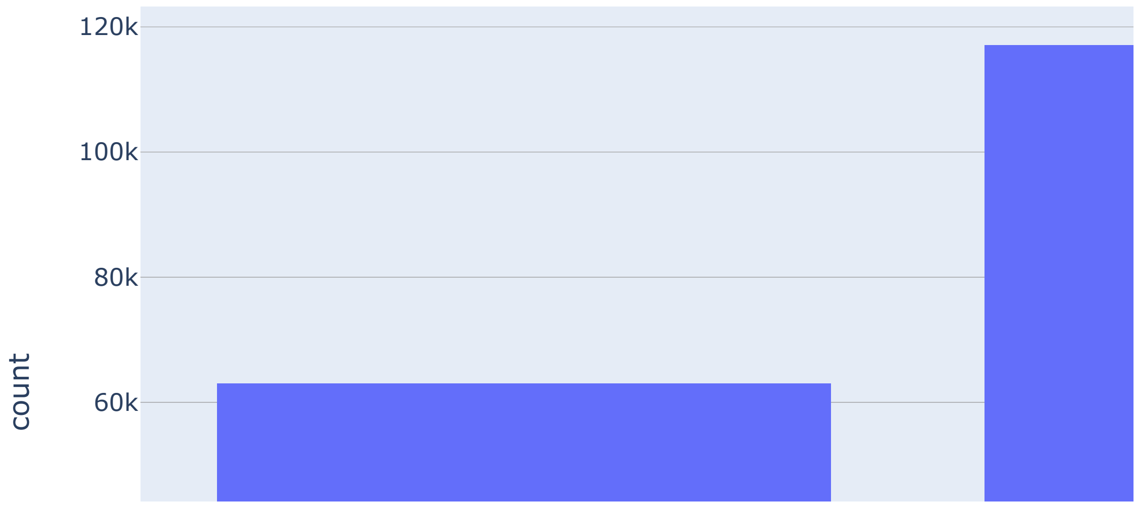
In [10]:

```
a = dataset['SEVERITY'].value_counts()
xx = a.index
yy = a.values
fig = px.bar(dataset['SEVERITY'], x=xx, y=yy, color = xx)
fig.show()
```



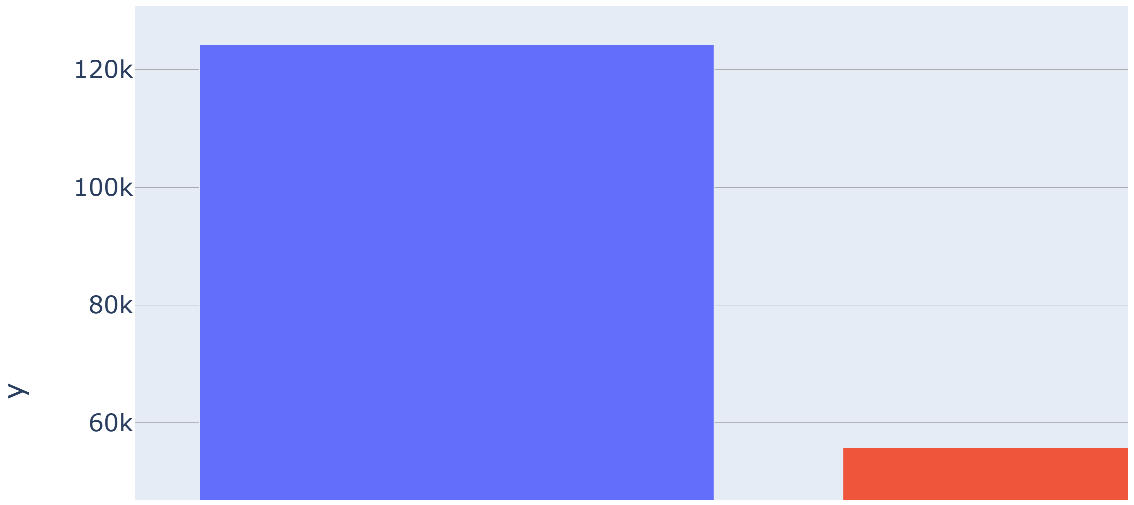
In [11]:

```
fig = px.histogram(dataset['ADDRTYPE'])
fig.show()
```



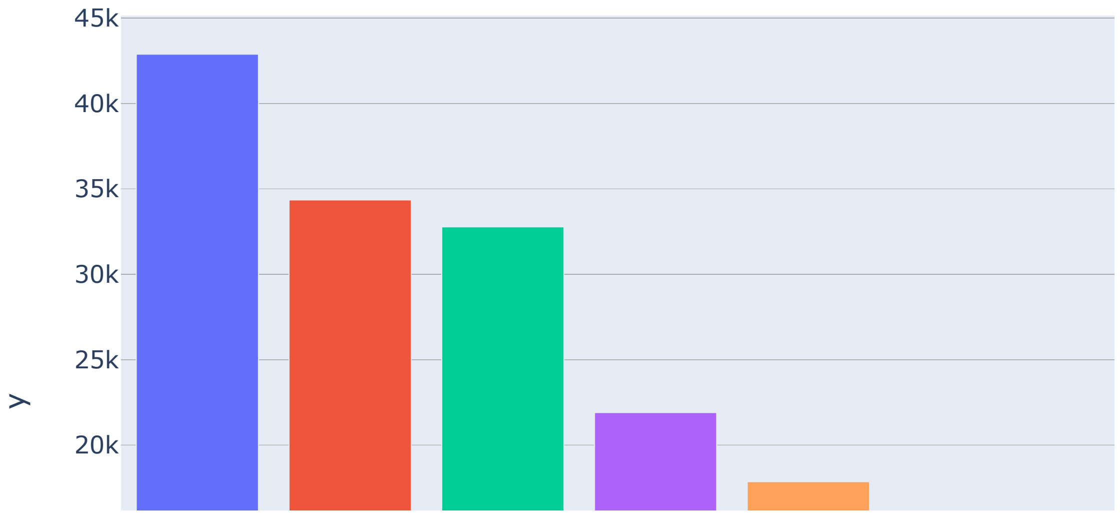
In [12]:

```
a = dataset['SEVERITYDESC'].value_counts()
xx = a.index
yy = a.values
fig = px.bar(dataset['SEVERITYDESC'], x=xx, y=yy, color = xx)
fig.show()
```



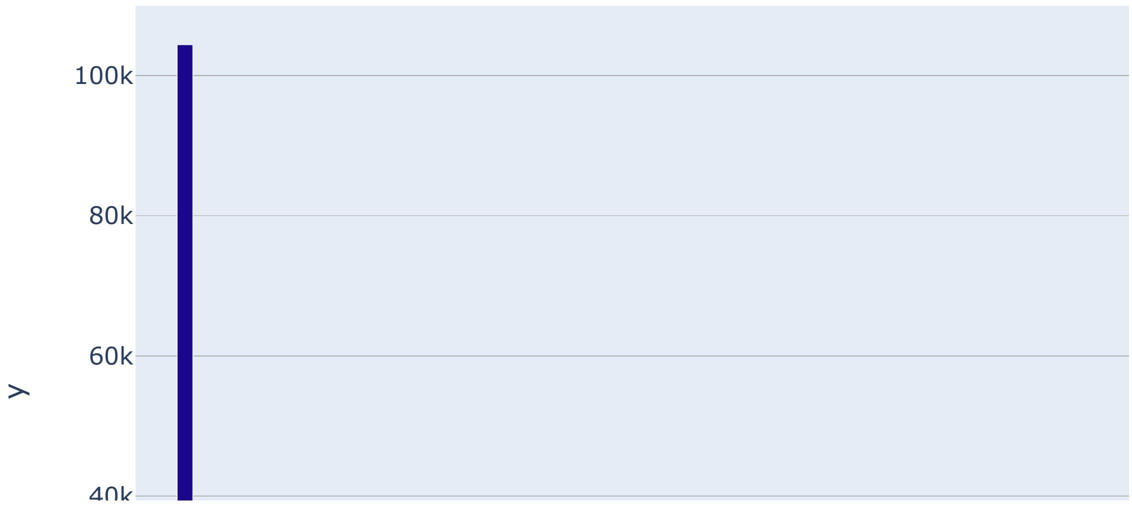
In [13]:

```
a = dataset['COLLISIONTYPE'].value_counts()
df = dataset['COLLISIONTYPE']
xx = a.index
yy = a.values
fig = px.bar(df, x=xx, y=yy, color = xx)
fig.show()
```



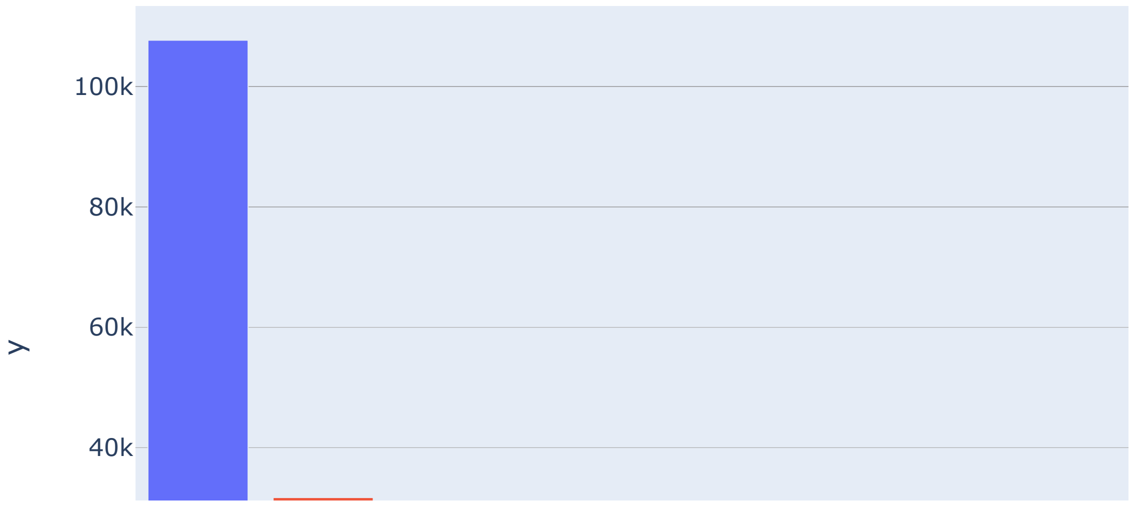
In [14]:

```
a = dataset['PERSONCOUNT'].value_counts()
df = dataset['PERSONCOUNT']
xx = a.index
yy = a.values
fig = px.bar(df, x=xx, y=yy, color = xx)
fig.show()
```



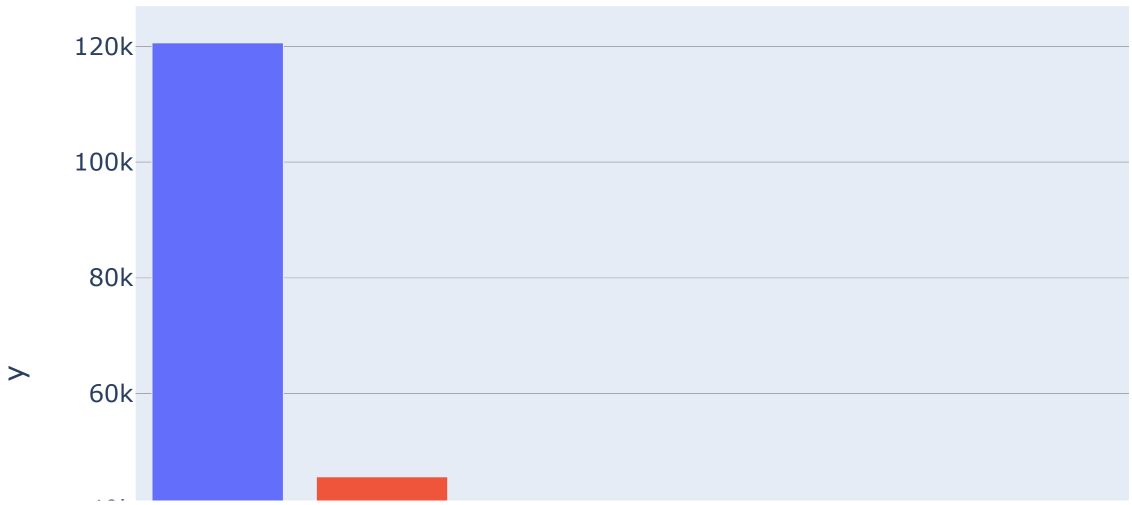
In [15]:

```
a = dataset['WEATHER'].value_counts()
df = dataset['WEATHER']
xx = a.index
yy = a.values
fig = px.bar(df, x=xx, y=yy, color = xx)
fig.show()
```



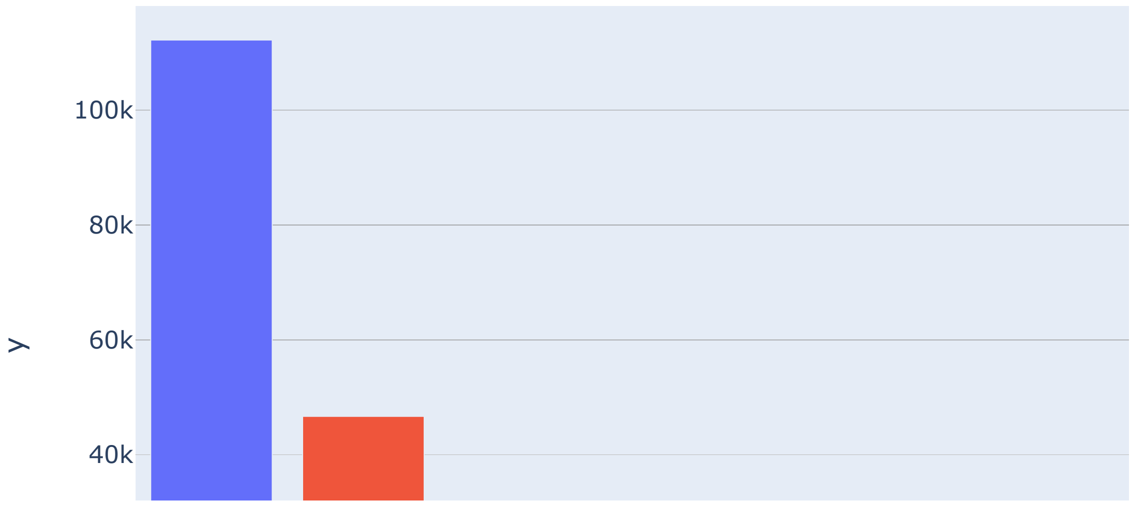
In [16]:

```
a = dataset['ROADCOND'].value_counts()
df = dataset['ROADCOND']
xx = a.index
yy = a.values
fig = px.bar(df, x=xx, y=yy, color = xx)
fig.show()
```



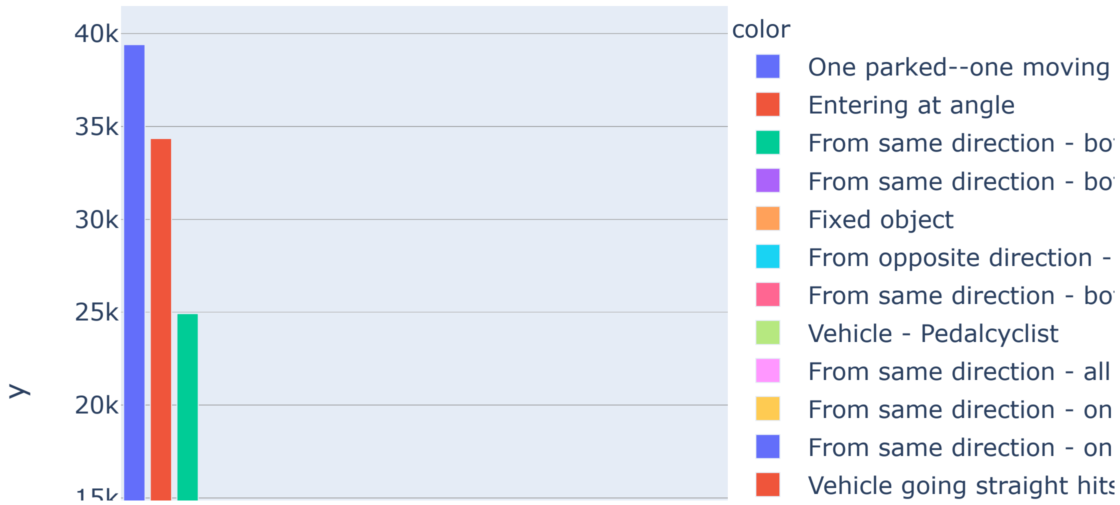
In [17]:

```
a = dataset['LIGHTCOND'].value_counts()
df = dataset['LIGHTCOND']
xx = a.index
yy = a.values
fig = px.bar(df, x=xx, y=yy, color = xx)
fig.show()
```



In [18]:

```
a = dataset['ST_COLDESC'].value_counts()
df = dataset['ST_COLDESC']
xx = a.index
yy = a.values
fig = px.bar(df, x=xx, y=yy, color = xx)
fig.show()
```



Following are the observations we have concluded after visualizing our data:

```
1. Number of accidents with Severity 1 is greater that that of Severity 2. Sever
ity 1 has total a of 124.258k fatalities while that of Severity 2 is of 55.809k.
2. More accidents occur at Blocks compared to Intersections. Number of accidents
at occured at Block are 117.085k while that of Intersections are 62.982k.
3. As seen in point No.1, it is good to see that most collisions caused only pro
perty damage like roads, vehicles etc. rather than causing Injuries. The numbers
are also the same - Property Damage = 124.258k & Injury Collisions = 55.809k
4. Top 3 accidents have occurred when:
        1 - Cars were parked and not moving. Total of 42.886k Fatalities.
        2 - At road angles.Probably occurred when one or more person(s) failed t
o notice another vehicle coming out from the                 other side of the r
oad. Total of 34.353k Fatalities.
        3 - At Rear Ends. This one occurs mostly when a person tries to overtake
another vehicle in front of them.
            Total of 32.778k Fatalities.
5. Maximum 2 to 3 Persons were involved in a particular accident. No of accident
s with 2 persons - 104.408k.  No of                accidents with 3 persons - 3
4.356k.
```

Most surprising thing to see is that most accidents have occured when one of the two cars involved in an accident was parked and still. Also most accidents have occured in broad Daylight when the weather conditions were good. This is probably because of the roads. The roads are not well maintained and must be crooky or bumped here and there. From my observation, it is the Roads that need maintenance although we will come to our conclusion only after applying our Machine Learning Models on this data.