## MISSION NAME :- Working with missing data

* Reading and counting null values in each column :-

```
import pandas as pd
mvc= pd.read_csv("nypd_mvc_2018.csv")
null_count= mvc.isnull().sum()
```

* The technical name for filling in a missing value with a replacement value is called imputation.

```
* Killed_cols = [col for col in mvc.columns if "killed" in col]
Killed = mvc[Killed_cols].copy()
killed_manual_sum = killed.iloc[:, :3].sum(axis=1)
killed_mask = killed_manual_sum != Killed["total_killed"]
killed_non_eq = killed[killed_mask]
```

* Series.mask(bool_mask, val_to_replace)

* Replacing missing values and cleaning suspicious data :-

```
injured = mvc[[col for col in mvc.columns if 'injured' in col]].
                                                            copy()
injured_manual_sum = injured.iloc[:, :3].sum(axis=1)
injured["total_injured"] = injured["total_injured"].mask(
                    injured["total_injured"].isnull(),
                    injured_manual_sum)
injured["total_injured"] = injured["total_injured"].mask(
                    injured["total_injured"] !=
                    injured_manual_sum,
                         np.nan)
```

* we can calculate the relationship between two sets of columns known as correlation.

* correlation function:- DataFrames.corr()

* np.triu(dataframe/matrix, k)

diagonal

Selects ~~Makes~~ the upper triangle ~~&~~ and elements below kth diagonal zero.

* np.ones_like(dataframe/matrix)

Makes a matrix of same shape as of dataframe with all values 1.

* To avoid removing rows with missing data, we can replace the null values with the values that appears most commonly.

* for that we can first bring the dataframe into a single series using DF.stack() method.

* Then on the series, applies Series.value_counts() method to find the most common one.

Eg:-

```
v_cols=[c for c in mvc.columns if c.startswith("vehicle")]
df = mvc[v_cols]
df_1d=df.stack()
top_10_vehicles= df_1d.value_counts().head(10)
```

* Replacing:-

```
for v in range(1,6):

    v_col = "vehide_{}".format(v)
    c_col = "cause_vehicle_{}".format(v)

    v_missing-mask= mvc [v_col]. isnull()& mvc [c_col].notnull()
    c_missing-mask= mvc [vcol].notnull() & mvc [c_col].isnull()

    mvc[v_col]= mvc[v_col].mask(v_missing-mask, "Unspecified")
    mvc [c_col]= mvc[c_col].mask(c_~~cod~~missing_mask, "Unspecified")
```