# install packages if needed  
# import pandas, numpy, matplotlib, seaborn, scikitlearn, folium, ...  
# import the following:  
'''  
pandas, numpy, matplotlib, seaborn,folium, standardscaler, minmaxscaler,  
chi2\_contingency, kmeans, pca, decision tree, plot tree, export\_graphviz,  
graphviz Source,tran test split,  
classification report, confusion matrix, accuracy score, transacation encoder,  
apriori, association rules,  
IPython display  
'''  
# notebook setup and config  
# mount google drive if you load/save data to/from your google drive  
# load the given full dataset  
# check shape, samples  
# select your subset based on the borough code  
# check shape samples  
# check data info  
# confirm borough  
# remove columns that you do not need (will mostly depend on your DM goals)  
# check data statistical summary  
# check data types - pandas  
# check data types - python  
# convert data types 1 - Date: string to datetime  
# convert data types 2 - Time: string to datetime  
# convert data types 3 - Police\_Force: numeric to categorical  
# convert data types 4 - Accident\_Severity: numeric to categorical-ordinal  
# convert data types 5 - Remaining columns - numeric to categorical as needed  
# check statistical summary - numeric columns  
# check statistical summary - object and cagtegory columns  
# check missing values - Note: some columns might have -1 as missing values  
# check distributions - numeric columns  
# check distributions - categorical columns  
# check geographical distribution of accidents - on a map  
# detect outliers - numeric columns  
# detect outliers - categorical columns [use value counts, domain knowledge, visualisations]  
# remove rows with missing values (or impute)  
# remove outliers (or impute)  
# create a Month column from Date column  
# create a Hour column from Time column  
# transform value ranges if needed - standardisation/normalisation  
# check pairwise correlation - numeric columns (all or any selected)  
# visualise correlations/relations - numeric columns  
# check correlations - categorical  
# visualise correlations - categorical  
# check correlations - categorical and numerical  
# visualise (co)relations - categorical and numerical  
# Kmeans clustering - column-set 1  
# cluster analysis for column-set 1  
# Kmeans clustering - column-set 2  
# cluster analysis for column-set 2  
# Association Rule Mining - generate rules  
# Association Rule Mining - analyse rules  
# decision tree - column set 1 - target 1  
# decision tree evaluation - column set 1 - target 1  
# decision tree visualisation - column set 1 - target 1  
# decision tree - column set 2 - target 2  
# decision tree evaluation - column set 2 - target 2  
# decision tree visualisation - column set 1 - target 1  
# neural network - column set 1 - target 1  
# neural network evaluation - column set 1 - target 1  
# neural network - column set 2 - target 2  
# neural network evaluation - column set 2 - target 2