SCS_3253 MACHINE LEARNING

TERM PROJECT

SCHOOL SAFETY ZONE: WATCH YOUR SPEED PROGRAM

Group 6
Aniket Ajit Ingale
Zlata Izvalava

WATCH YOUR SPEED PROGRAM

- The Watch Your Speed Program (WYSP) uses devices called speed display signs or driver feedback signs which contain a radar device and an LED display.
- The radar measures the speeds of oncoming vehicles and the LED sign displays their speeds to the passing motorists, thereby reminding them to check their speeds and to obey speed limits. The City of Toronto's permanent units are installed in Safety Zones.
- We used the data recorded by permanent driver feedback signs installed in School Safety Zones. The datasets are published by Transportation Services on the City of Toronto Open Data Portal.

DATA

1. **"School Safety Zone Watch Your Speed Program – Detailed Speed Counts"**
   An hourly aggregation of observed speeds for each location where a Watch Your Speed Program Sign was installed in 10 km/hr speed range increments.
   https://open.toronto.ca/dataset/school-safety-zone-watch-your-speed-program-detailed-speed-counts/
2. **"School Safety Zone Watch Your Speed Program – Locations"**
   The locations and operating parameters for each location where a permanent Watch Your Speed Program Sign was installed.
   https://open.toronto.ca/dataset/school-safety-zone-watch-your-speed-program-locations/

Limitations:

- The count of number of vehicles is not equivalent to a traffic volume count
- Sign addresses have not been verified

PROBLEM

1. Explore the data recorded by the driver feedback sign #230 (approximate address - 994 Jane Street, Toronto;
   southbound direction of travel;
   speed limit - 50 km/hr)
2. Train a model to predict an hourly count of vehicles traveling at a speed in the "50 km/hr and higher" speed range for the location where the sign #230 was installed.

Analysis of recorded data and the model can be useful for understanding the situation with safety in this school zone and for planning measures to improve it.

Note: The model cannot be used during the pandemic because the situation in the city has changed (traffic volume has reduced drastically, schools are closed, etc.)

Approximate location of the WYSP sign #230

APPROACH

1. Download the datasets and read them into pandas DataFrames
2. Prepare the data (combine all records for the sign #230, reshape the DataFrame and add new columns)
3. Explore the data
4. Create training and test sets, transform categorical attributes
5. Train different regression models and compare the results
6. Evaluate 2 best models on the test set and choose the final model
7. Create a pipeline for final model and use it for prediction example

LOAD THE DATA

1. **"School Safety Zone Watch Your Speed Program – Detailed Speed Counts"**
   Stationary count detailed 2019 (zip): includes 12 csv files – for all months in 2019
   Stationary count detailed 2020 (zip): includes 2 csv files – for Jan and Feb 2020
   Stationary Detail Counts Readme (xlsx): contains column descriptions for datasets

2. **"School Safety Zone Watch Your Speed Program – Locations"**
   Stationary Sign locations (csv): details on where and when the signs were installed.
   Information about the sign #230:
   Data was recorded from 2019-03-20 06:00 to 2020-02-19 23:00
   Speed limit – 50 km/hr
   "schedule": Weekdays from 7 AM - 9 PM (Times of week when the sign is on. Signs still record speeds when the display is inactive)

The data will be downloaded, extracted and read into pandas DataFrames when you run the Notebook 1.

PREPARE THE DATA

1. **Combine all records for the sign #230 into one DataFrame** (no missing values)
   Period from March 2019 to February 2020 - when data was recorded by this sign

	sign_id	address	dir	datetime_bin	speed_bin	volume
696125	230	994 Jane Street	SB	2019-03-20T06:00	[0,10)	7
696126	230	994 Jane Street	SB	2019-03-20T06:00	[10,20)	11
696127	230	994 Jane Street	SB	2019-03-20T06:00	[20,30)	16
696128	230	994 Jane Street	SB	2019-03-20T06:00	[30,40)	47
696129	230	994 Jane Street	SB	2019-03-20T06:00	[40,50)	99

"datetime_bin": Start of the hour during which these observations were made

"speed_bin": Range of speeds observed (e.g. [10,20) represents speeds from 10 km/hr up to and not including 20 km/hr)

"volume": Number of vehicles observed in that hour and speed bin

PREPARE THE DATA

2. **Reshape the DataFrame from "long" to "wide" format**
   Initially, there were a few rows for each timestamp that represented hourly speed counts for different speed bins. After reshaping there is one row for each timestamp with all observations.
   Created missing values were replaced with zeros (because no vehicles were observed in that speed range during that hour)

speed_bin	[0,10)	[10,20)	[100,)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100)
datetime_bin											
2019-03-20T06:00	7	11	0	16	47	99	103	43	4	2	0
2019-03-20T07:00	0	1	0	6	10	36	37	15	3	0	0
2019-03-20T08:00	1	9	0	18	44	106	110	37	4	2	0
2019-03-20T09:00	2	10	0	16	42	136	182	89	12	2	0
2019-03-20T10:00	4	4	0	9	24	96	162	97	14	0	0

PREPARE THE DATA

3. **Create new columns for analysis and modelling**
   - 'hour' column: values from 0 to 23
   - 'day_of_week' column: values from 0 (Sunday) to 6
   - 'month' column: values from 1 to 12
   - 'display_on' column: values 1 (on) and 0 (off)
     Display schedule for sign #230 when display is active: Weekdays from 7AM to 9PM
   - 'total_count' column: total number of vehicles observed during that hour
   - 'over_50' column (the target column): number of vehicles observed traveling at speeds of 50 km/hr and higher during that hour
   - 'percent_over_50' column: proportion of vehicles traveling at speeds in the "50 km/hr and higher" speed range during certain hour

EXPLORE THE DATA

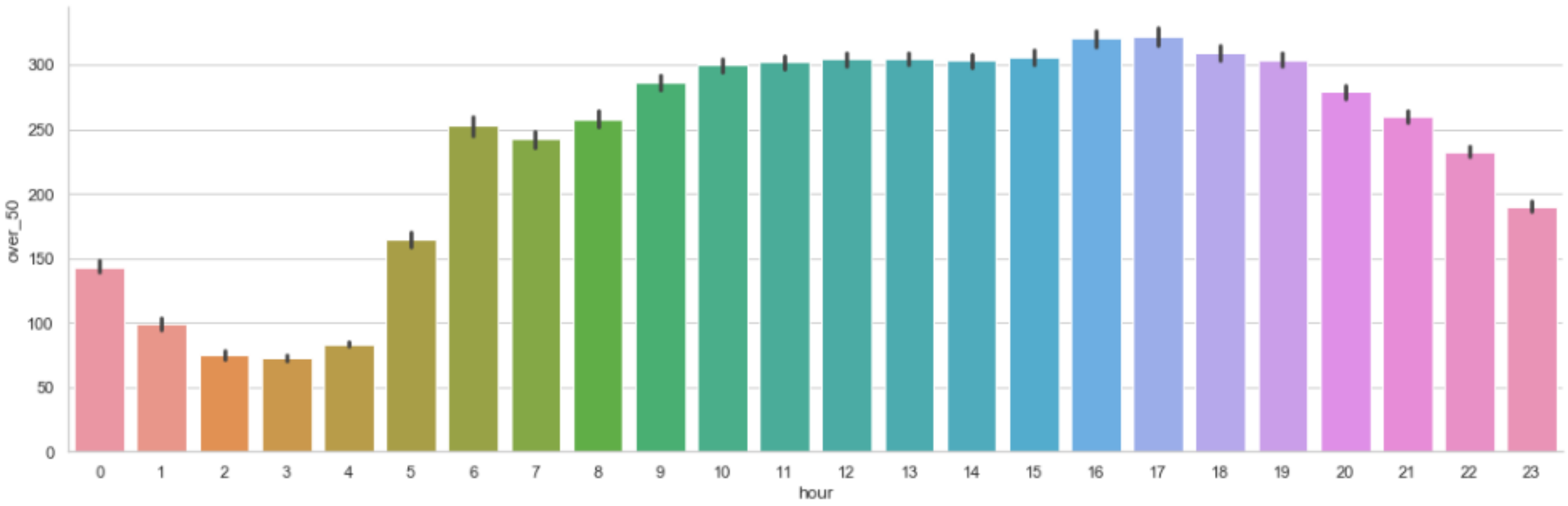
Information about the DataFrame:

- Total number of rows = 7973. Each row represents one timestamp - start of the hour during which the observations were made
- There are no missing values
- All columns have numerical values (but the 'month', 'day_of_week', 'hour' and 'display_on' columns are categorical attributes)

Summary statistics:

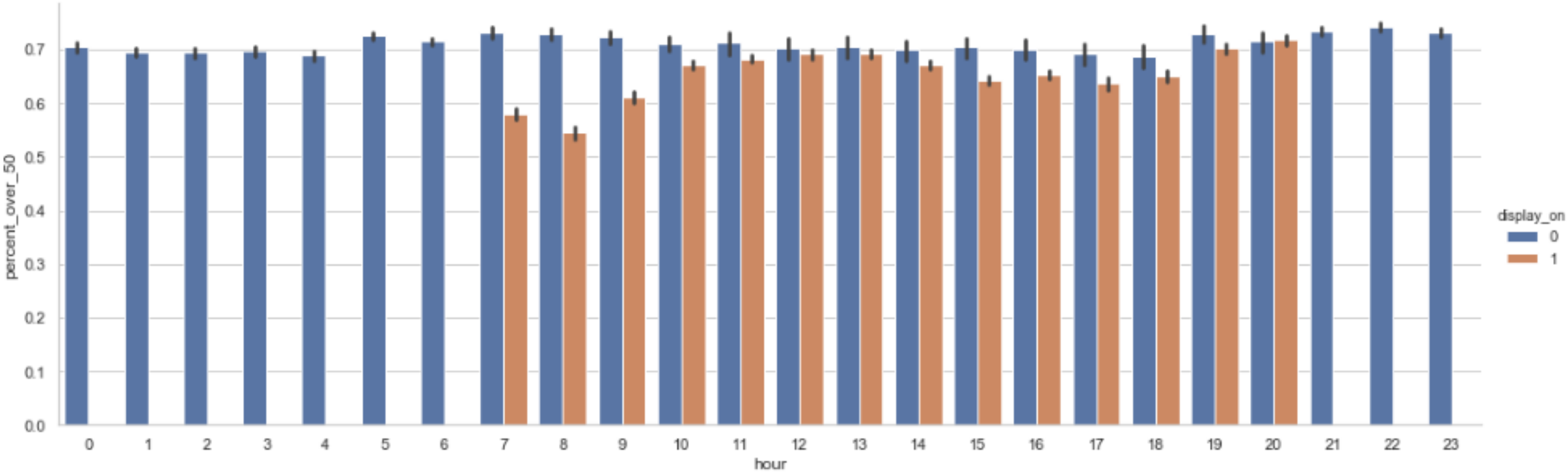
	month	day_of_week	hour	display_on	total_count	over_50	percent_over_50
count	7973.000000	7973.000000	7973.000000	7973.000000	7973.000000	7973.000000	7973.000000
mean	6.858021	3.001003	11.486141	0.415904	349.445127	237.997868	0.687641
std	3.405289	2.005762	6.921970	0.492908	141.868034	96.355271	0.089116
min	1.000000	0.000000	0.000000	0.000000	5.000000	2.000000	0.020548
25%	4.000000	1.000000	5.000000	0.000000	239.000000	168.000000	0.647799
50%	7.000000	3.000000	11.000000	0.000000	400.000000	266.000000	0.701847
75%	10.000000	5.000000	17.000000	1.000000	461.000000	311.000000	0.745562
max	12.000000	6.000000	23.000000	1.000000	719.000000	486.000000	0.896226

Group the ‘over_50’ data by hour (compute group means)



The highest values are observed for hours 4pm-5pm and 5pm-6pm

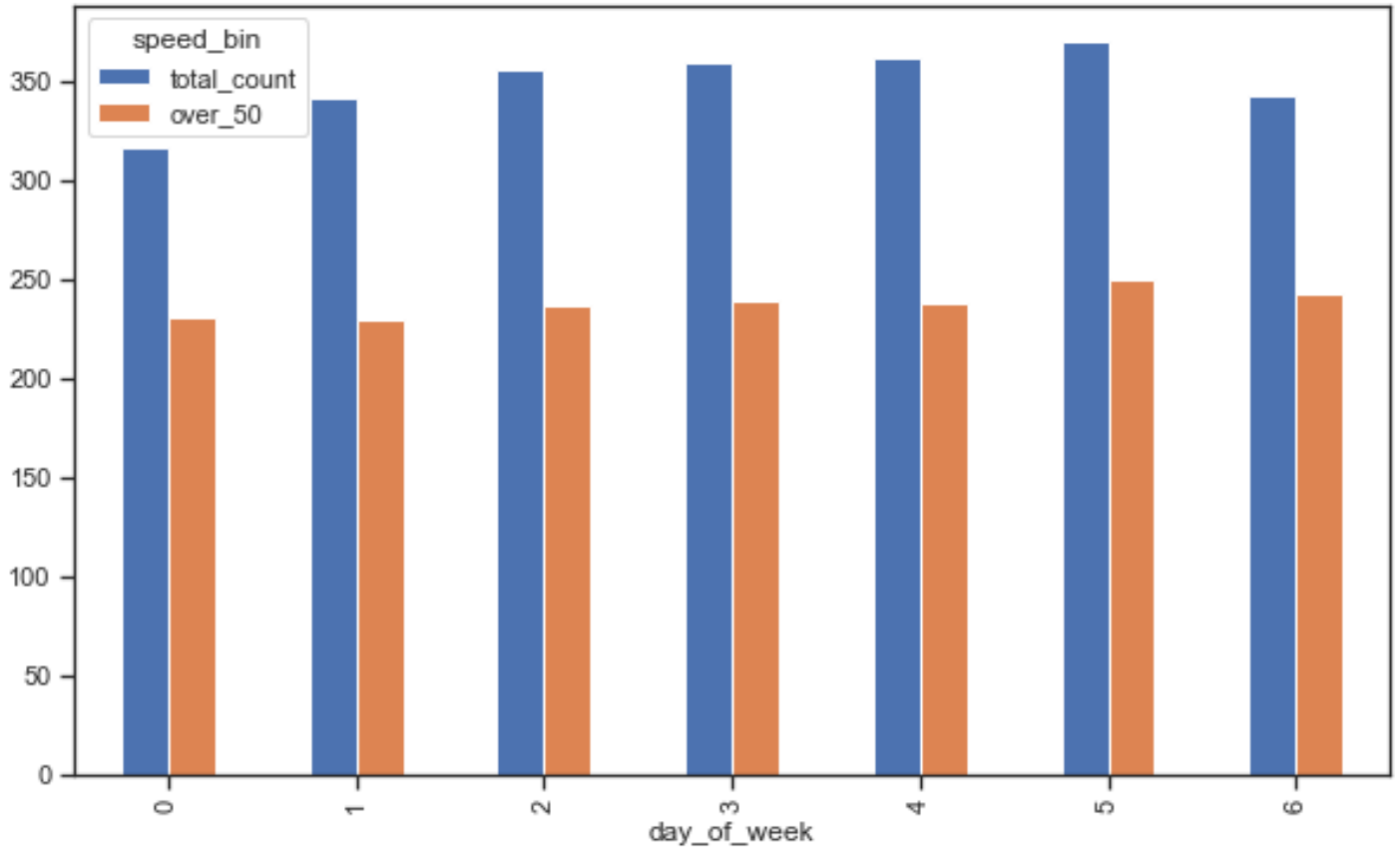
Proportion of vehicles in the “50 km/hr and higher” speed bin grouped by hour and by display on/off (compute group means)



The average percentage is mostly higher when display is off than when it is active

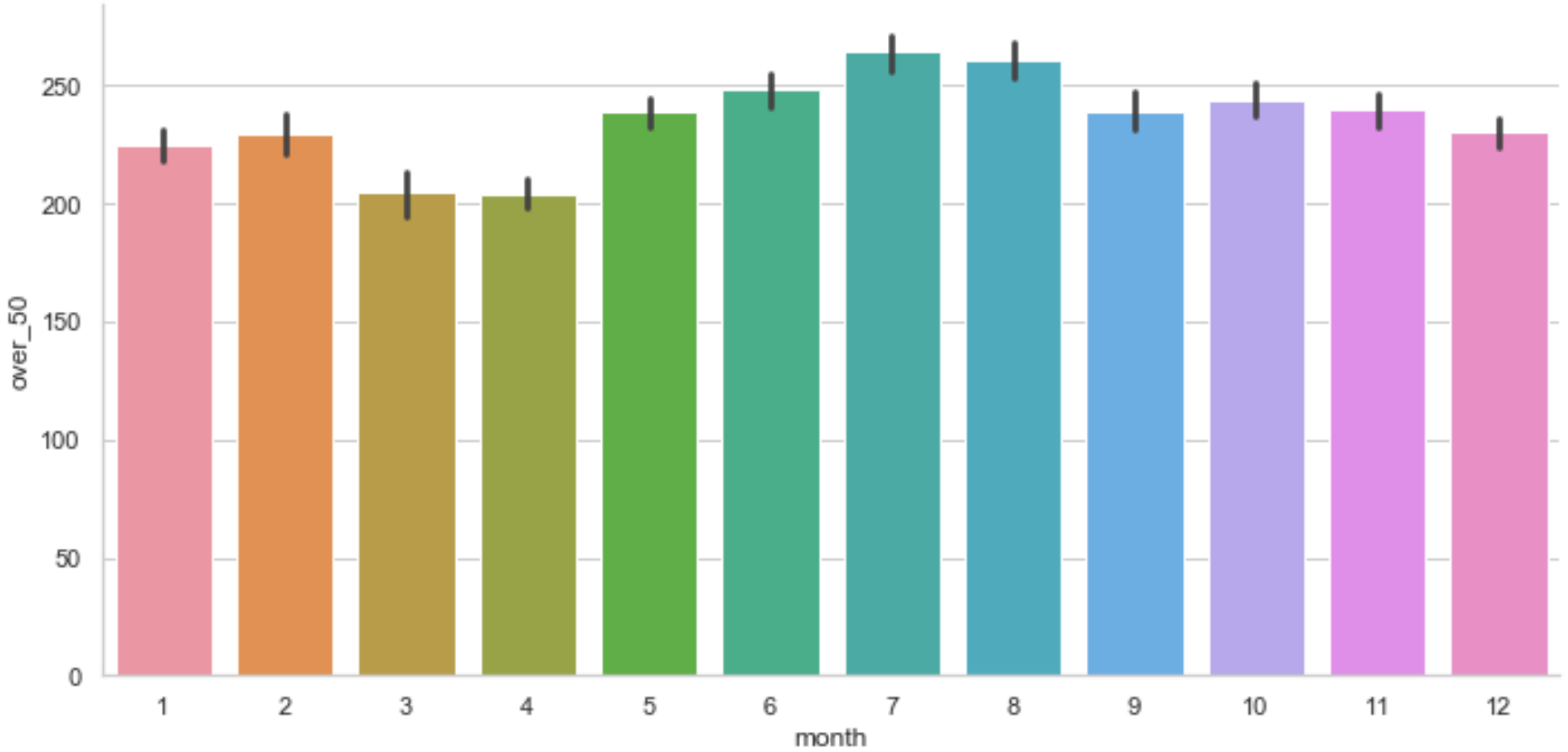
Group the data by day of week (compute group means)

day_of_week	total_count	over_50	percent_over_50
0	316.364111	230.618467	0.730317
1	340.864629	229.231441	0.679066
2	356.285333	236.736889	0.670607
3	359.585106	238.915780	0.671883
4	361.502210	238.068966	0.666062
5	369.480903	249.506944	0.681557
6	342.409091	242.852273	0.713146



After grouping, the highest 'over_50' value is observed for Friday and the highest 'percent_over_50' value - for Sunday

Group the ‘over_50’ data by month (compute group means)



The highest values are observed for July and August – months when the school is closed

TRANSFORM THE DATA FOR MODELLING

1. Create a DataFrame containing the features and the target.
   Features: 'hour', 'day_of_week', 'month'
   Target: 'over_50'
2. Create training and test sets: stratified split based on the 'month' column to keep the category proportions (less data was recorded for Mar 2019 and Feb 2020).
   Test size = 0.2
3. Transform categorical attributes using OneHotEncoder:
   43 features after transformation

TRAINING REGRESSION MODELS

Model	Hyperparameters	RMSE (training set)	Mean RMSE (cv=3)
Linear Regression		43.90	44.24
Lasso Regression	alpha = 0.001	43.90	44.25
Ridge Regression	alpha = 0.1	43.90	44.25
Polynomial Regression	**degree = 2**	**33.94**	**37.76**
Linear SVM Regression		45.61	47.47
SVM Regression using a second-degree polynomial kernel	kernel = "poly", degree = 2	71.07	79.70
XGBoost Regression	gamma = 0.8	42.60	43.48
Random Forest Regression	Standard (before tuning)	29.64	41.11
	n_estimators = 200, min_samples_leaf = 8	**35.13**	**38.15**

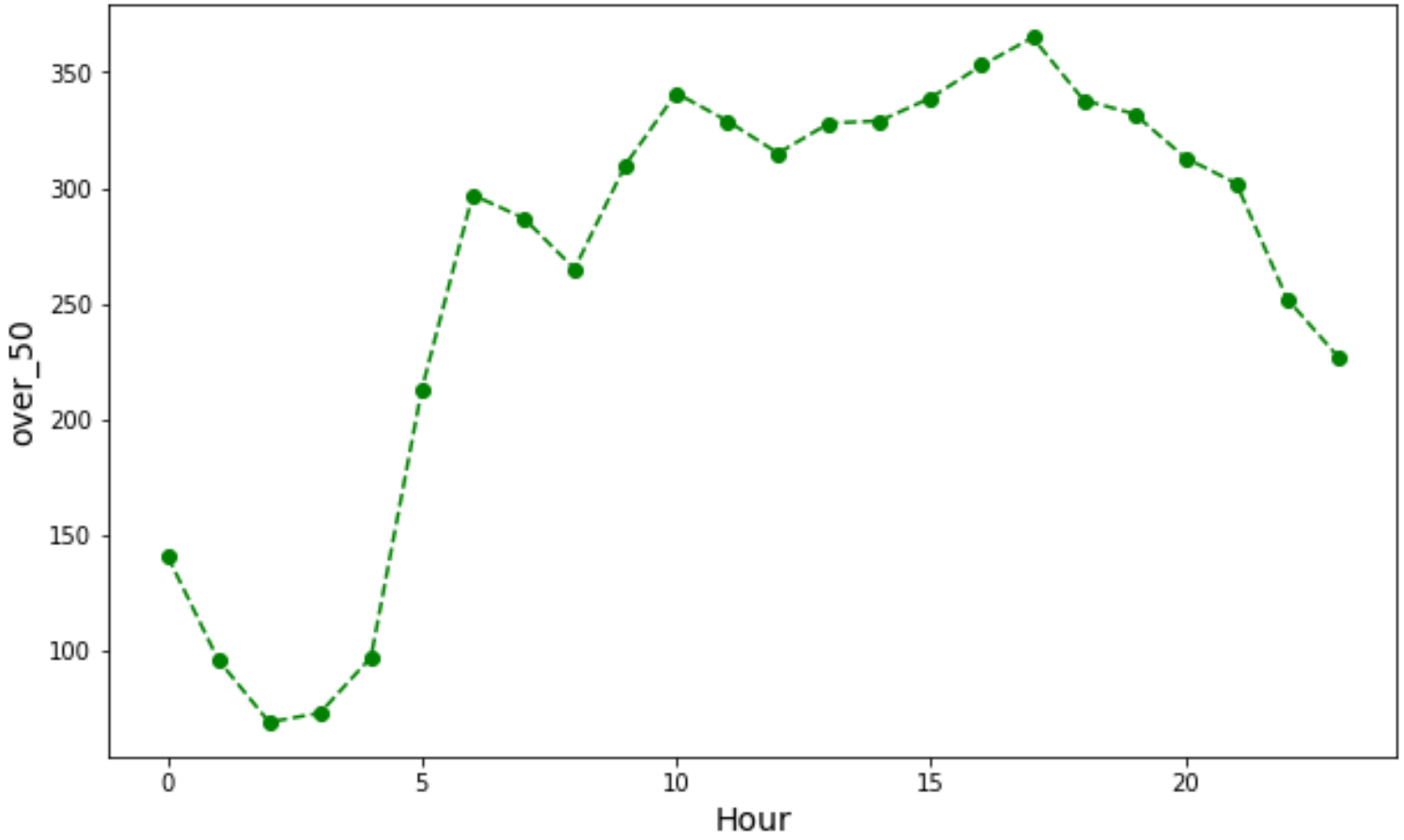
EVALUATE 2 BEST MODELS ON THE TEST SET

Model	Hyperparameters	RMSE (training set)	Mean RMSE (cv=3)	RMSE (test set)	95% Confidence interval for the test RMSE
Polynomial Regression	degree = 2	33.94	37.76	36.97	[33.86, 39.83]
Random Forest Regression	n_estimators = 200, min_samples_leaf = 8	35.13	38.15	37.65	[34.58, 40.49]

The results for two models are very close. Polynomial Regression model performed slightly better, it was chosen as a final model.

The estimates on the test set are off by an average of 15.6% for Polynomial Regression model.

Example: Using the final model to make predictions for October 16, 2020 (if city traffic is back to normal)



The highest ‘over_50’ count is expected between 5 pm and 6 pm (365 vehicles)

CONCLUSION

- Analysis of the data recorded by the Watch Your Speed Program sign #230 showed that many drivers did not obey speed limits in this location (time period from 2019-03-20 to 2020-02-19).
- The use of the Polynomial Regression model (degree = 2) can help the City of Toronto to plan and take measures to improve safety in this school safety zone.
- Similar analysis can be done for other locations where permanent driver feedback signs were installed.

THANK YOU!