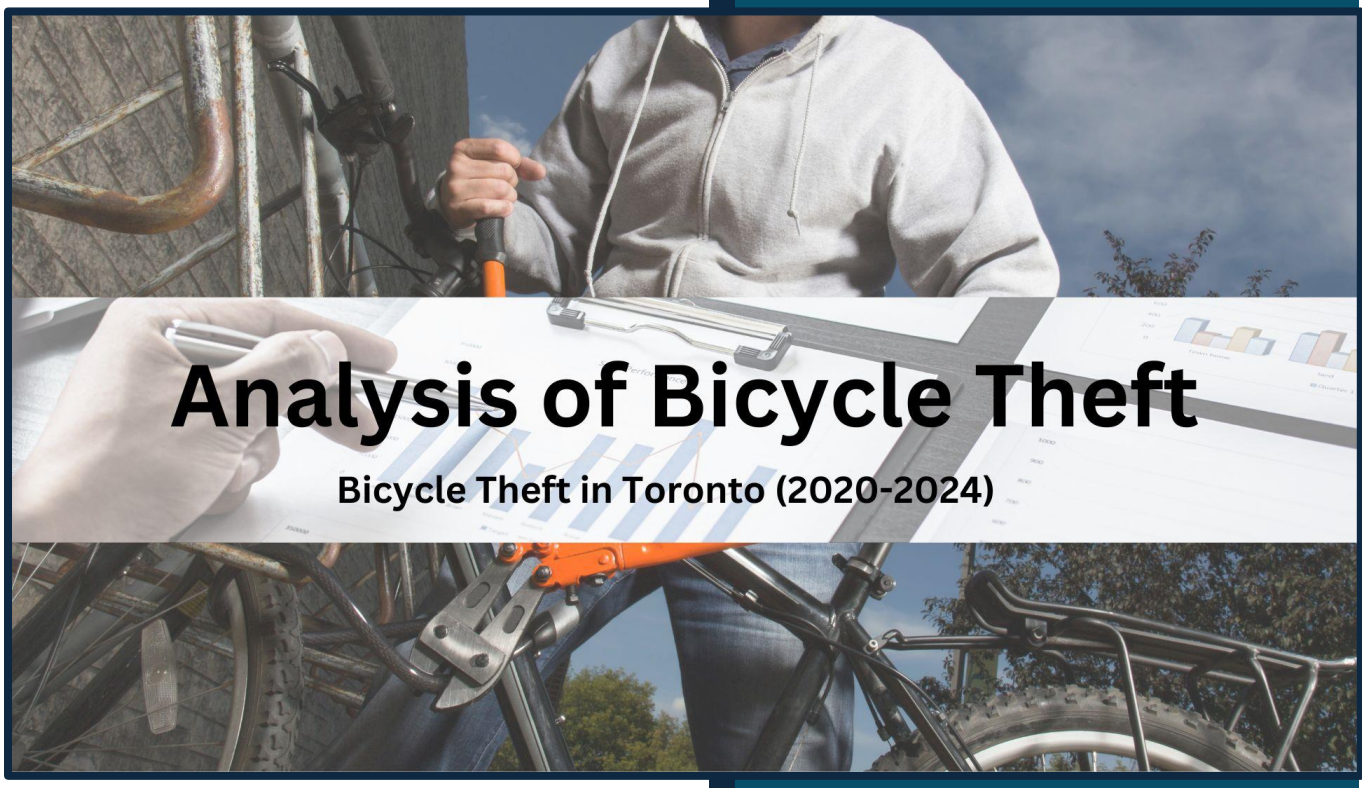
2024



Analysis of Bicycle Theft

Bicycle Theft in Toronto (2020-2024)

Mateus Augusto Ali Fontes

Metro College of Technology

12/13/2024

Contents

Summary 2
About dataset 3
Where I find this dataset? 3
Technical information about Bicycle Thefts dataset. 4
Contents about dataset. 4
Segmentations / Descriptions 6
Univariate Analysis 8
Q1 - Annual Bicycle Theft Analysis 8
Q2 - Monthly Bicycle Theft 9
Q3 - During which period of the day are bicycles most frequently stolen 11
Q4 - Bicycle Theft by Day of the Week 12
Q5 - Bicycle Theft Type 14
Q6 - Bicycle Theft by Premise Type 16
Q7 - What is the average price of a stolen bicycle? 18
Q8 - What is the average maximum speed of stolen bicycles? 20
Q9 - How many bicycles were recovered? 21
Q10 - How many days, on average, pass between the date a bicycle is stolen and the date the theft is reported? 22
Bivariate Analysis 23
Q1 - What has been the growth in e-bike thefts from 2020 to 2024? 23
Q2 - Does the day of the week affect the cost of the bike that was stolen? 24
Q3 - Does the time of day affect the cost of the bike that was stolen? 26
Q4 - Is there any relationship between the premises where the bike was stolen and the season of the year? 28
Q5 - Is there any correlation between the maximum speed of the bike and its cost? 30
Q6 - Is there any relationship between the average bike cost and weekdays versus weekends? 32
Q7 – 35
Conclusion 37

Summary

In this report you can find some analysis about dataset that contains information's about bicycles was theft in Toronto – Canada for the last 5 years.

I performed some univariate and bivariate analysis that is possible we have some conclusions about the data, and make some predictions.

About dataset

Where I find this dataset?

This dataset contains Bicycle Thefts occurrences by reported date and details regarding the stolen item where available. This data includes all bicycle theft occurrences reported to the Toronto Police Service, including those where the location has not been able to be verified. As a result, coordinate fields may appear blank. Likewise, this includes occurrences where the coordinate location is outside the City of Toronto.

Note: Fields have been included for both the old 140 City of Toronto Neighbourhoods structure as well as the new 158 City of Toronto Neighbourhoods structure

Limitations

This dataset contains Bicycle Thefts occurrences from 2014-2024/09. The location of crime occurrences has been deliberately offset to the nearest road intersection node to protect the privacy of parties involved in the occurrence. Due to the offset of occurrence location, the numbers by Division and Neighbourhood may not reflect the exact count of occurrences reported within these geographies. Therefore, the Toronto Police Service does not guarantee the accuracy, completeness, timeliness of the data and it should not be compared to any other source of crime data.

Source: https://open.toronto.ca/dataset/bicycle-thefts/

Technical information about Bicycle Thefts dataset.

This dataset was not normally, then I for better visualization and analysis I cleaned it, excluding all observations lower from first quartile and all observations upper from third quartile. (SAS code line 28 to 47).

Contents about dataset.

The CONTENTS Procedure

Data Set Name	ALI.BICYCLE_THEFTS_2020_2024	Observations	11363
Member Type	DATA	Variables	30
Engine	V9	Indexes	0
Created	2024-12-17 10:28:45	Observation Length	432
Last Modified	2024-12-17 10:28:45	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	76
First Data Page	1
Max Obs per Page	151
Obs in First Data Page	141
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	D:\DSA - College\FDA\bicycle_thefts_2020_2024.sas7bdat
Release Created	9.0401M7
Host Created	X64_10PRO
Owner Name	DESKTOP-C8GPL1G\mateu
File Size	5MB
File Size (bytes)	5046272

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
25	BIKE_COLOUR	Char	6	$6.	$6.
26	BIKE_COST	Num	8	BEST12.	BEST32.
21	BIKE_MAKE	Char	15	$15.	$15.
22	BIKE_MODEL	Char	17	$17.	$17.
24	BIKE_SPEED	Num	8	BEST12.	BEST32.
23	BIKE_TYPE	Char	2	$2.	$2.
18	DIVISION	Char	3	$3.	$3.
2	EVENT_UNIQUE_ID	Char	14	$14.	$14.
29	LAT_WGS84	Num	8	BEST12.	BEST32.
19	LOCATION_TYPE	Char	72	$72.	$72.
28	LONG_WGS84	Num	8	BEST12.	BEST32.
4	OCC_DATE	Num	8	YYMMDD10.	YYMMDD10.
8	OCC_DAY	Num	8	BEST12.	BEST32.
7	OCC_DOW	Char	9	$9.	$9.
9	OCC_DOY	Num	8	BEST12.	BEST32.
10	OCC_HOUR	Num	8	BEST12.	BEST32.
6	OCC_MONTH	Char	9	$9.	$9.
5	OCC_YEAR	Num	8	BEST12.	BEST32.
20	PREMISES_TYPE	Char	11	$11.	$11.
3	PRIMARY_OFFENCE	Char	30	$30.	$30.
11	REPORT_DATE	Num	8	YYMMDD10.	YYMMDD10.
15	REPORT_DAY	Num	8	BEST12.	BEST32.
14	REPORT_DOW	Char	9	$9.	$9.
16	REPORT_DOY	Num	8	BEST12.	BEST32.
17	REPORT_HOUR	Num	8	BEST12.	BEST32.
13	REPORT_MONTH	Char	8	$8.	$8.
12	REPORT_YEAR	Num	8	BEST12.	BEST32.
27	STATUS	Char	9	$9.	$9.
1	_id	Num	8	BEST12.	BEST32.
30	geometry	Char	92	$92.	$92.

For my analysis univariate and bivariate I just used this follow variables:

OCC_DOW	Day of the Week Offence Occurred
REPORT_DATE	Date Offence was Reported
OCC_DATE	Date of Offence
OCC_YEAR	Year Offence Occurred
BIKE_SPEED	Speed of Bicycle
BIKE_COST	Cost of Bicycle
STATUS	Status of Bicycle
PREMISES_TYPE	Premises Type of Offence
OCC_MONTH	Month Offence Occurred
OCC_HOUR	Hour Offence Occurred
BIKE_TYPE	Type of Bicycle

Segmentations / Descriptions

Segmentations created to analysis better the data set (SAS code lines 52 to 139)

For the variable BIKE_SPEED, I divided the observations into four groups:
- "Non-motorized" includes all observations below 1 km/h.
- "Up to 25 km/h" includes all observations from 1 km/h to 25 km/h.
- "Up to 50 km/h" includes all observations from 26 km/h to 50 km/h.
- "Above 50 km/h" includes all observations above 51 km/h.

For the variable BIKE_TYPE, I classified different types of bikes:
- BMX, Electric, Folding, Mountain, Road, Touring bikes, and others were labeled for clear identification.

For the variable BIKE_COST, I divided bike prices into ranges:
- "Below $500" includes bikes costing up to $500.
- "$500 - $1000" includes bikes in that price range.
- "$1000 - $2000" includes bikes priced between $1000 and $2000.
- Higher ranges continue up to "Above $5000."

For the variable DIF_DAYS, I segmented the number of days into periods:
- "Up to 5 Days" includes observations within 5 days.
- "6 to 10 Days" through "Over 50 Days" covers progressively longer time spans.

For the variable HOURPERIODS, I divided hours into four periods:
- "Night" from 0 to 6 hours.
- "Morning" from 7 to 12 hours.
- "Afternoon" from 13 to 18 hours.
- "Evening" from 19 to 23 hours.

For the variable SEASON, I categorized months into seasons:
- "Spring" includes March, April, and May.
- "Summer" includes June, July, and August.
- "Fall" includes September, October, and November.
- "Winter" includes December, January, and February.

For the variable WKDAYS, I classified days of the week into:
- "Weekday" for Monday through Friday.
- "Weekend" for Saturday and Sunday.

For the variable REPORTDAYS, I divided reporting times into:
- "On-time Report" for observations reported within 14 days.
- "Late Report" for reports submitted after 14 days.

Univariate Analysis

Q1 - Annual Bicycle Theft Analysis

To determine the number of bicycles stolen each year, we performed a frequency analysis using the PROC FREQ statement. This allowed us to calculate the frequency and percentage of bicycle theft incidents by year.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: OCC_YEAR

Visualization Method: Vertical Bar Chart (VBAR) / Box Plot

Analysis Technique: PROC FREQ

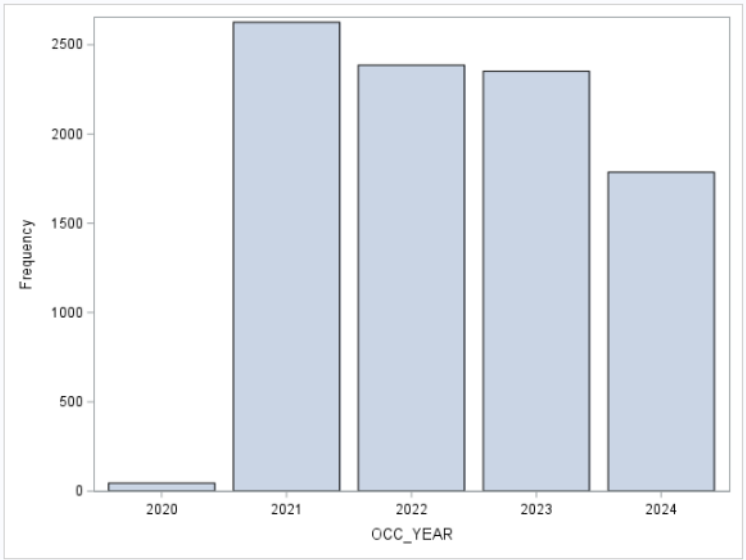
Univariate - Qualitative

Variable:OCC_YEAR (Year Offence Occurred)

Q1 - How many bicyles was thetf by year

The F REQ Procedure

OCC_YEAR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2020	45	0.49	45	0.49
2021	2626	28.56	2671	29.05
2022	2386	25.95	5057	55.00
2023	2351	25.57	7408	80.57
2024	1786	19.43	9194	100.00



Conclusion: As we can see, the number of bike thefts in Toronto has started to decrease from 2021 to 2024, dropping from 2,626 to 1,786.

Q2 - Monthly Bicycle Theft

To analyze the distribution of bicycle thefts across different months, we used the PROC FREQ statement to calculate the frequency and percentage of theft incidents by month.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: OCC_MONTH

Visualization Method: Vertical Bar Chart (VBAR)

Analysis Technique: PROC FREQ

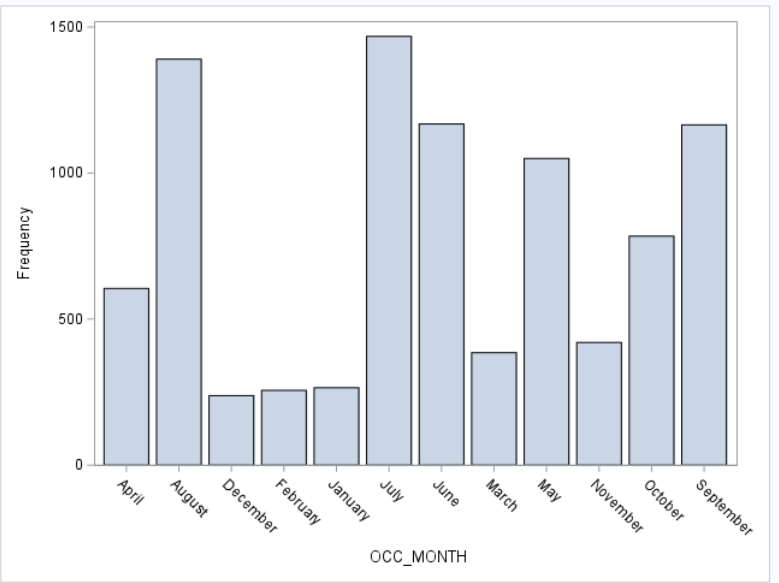
Univariate - Qualitative

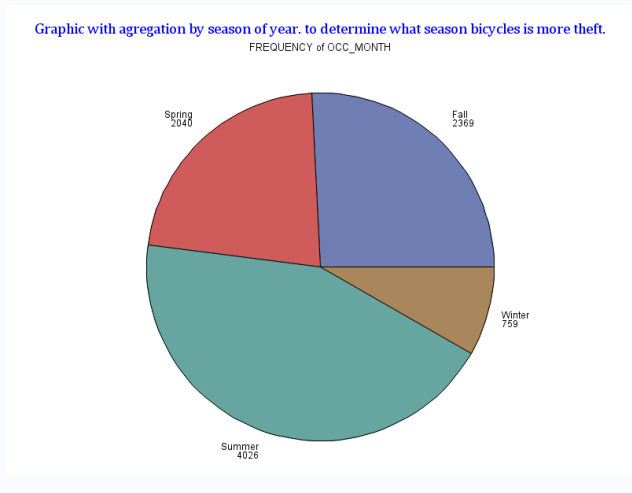
Variable:OCC_MONTH (Month Offence Occurred)

Q2 - How many bicyles was thetf by month ?

The FREQ Procedure

OCC_MONTH	Frequency	Percent	Cumulative Frequency	Cumulative Percent
July	1468	15.97	1468	15.97
August	1390	15.12	2858	31.09
June	1168	12.70	4026	43.79
September	1165	12.67	5191	56.46
May	1050	11.42	6241	67.88
October	784	8.53	7025	76.41
April	605	6.58	7630	82.99
November	420	4.57	8050	87.56
March	385	4.19	8435	91.74
January	265	2.88	8700	94.63
February	256	2.78	8956	97.41
December	238	2.59	9194	100.00





Conclusion:

In this analysis of the frequency of bicycle thefts by month, we observe that July is the month with the highest number of bike thefts. Additionally, we can infer that the season of the year directly affects the number of bike thefts. As shown in the pie chart, summer is the season with the most thefts.

Q3 - During which period of the day are bicycles most frequently stolen

To answer this questions we need use the statement PROC FREQ to check the frequencies and calculate percentage of how many bicycles was theft in specifically hour of the day, also we need do some aggregation. Check proc format *hourperiods*.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: OCC_HOUR

Visualization Method: Vertical Bar Chart (VBAR)

Analysis Technique: PROC FREQ

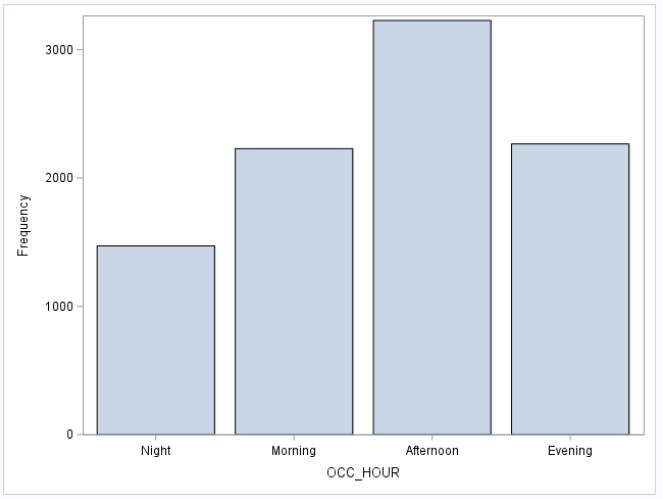
Univariate - Qualitative

Variable:OCC_HOUR (Hour Offence Occurred)

Q3 - During which period of the day are bicycles most frequently stolen?

The FREQ Procedure

OCC_HOUR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Night	1471	16.00	1471	16.00
Morning	2229	24.24	3700	40.24
Afternoon	3228	35.11	6928	75.35
Evening	2266	24.65	9194	100.00



Conclusion:

In this analysis of the **OCC_HOUR** variable, we observe that the afternoon is the time of day with the highest number of bike thefts. Additionally, I applied segmentation to the hours of the day, which can be reviewed on page X.

Q4 - Bicycle Theft by Day of the Week

To examine the distribution of bicycle theft incidents by day of the week, we used the PROC FREQ statement. This helped us calculate the frequency and percentage of theft occurrences for each day.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: OCC_DOW

Visualization Method: Vertical Bar Chart (VBAR)

Analysis Technique: PROC FREQ

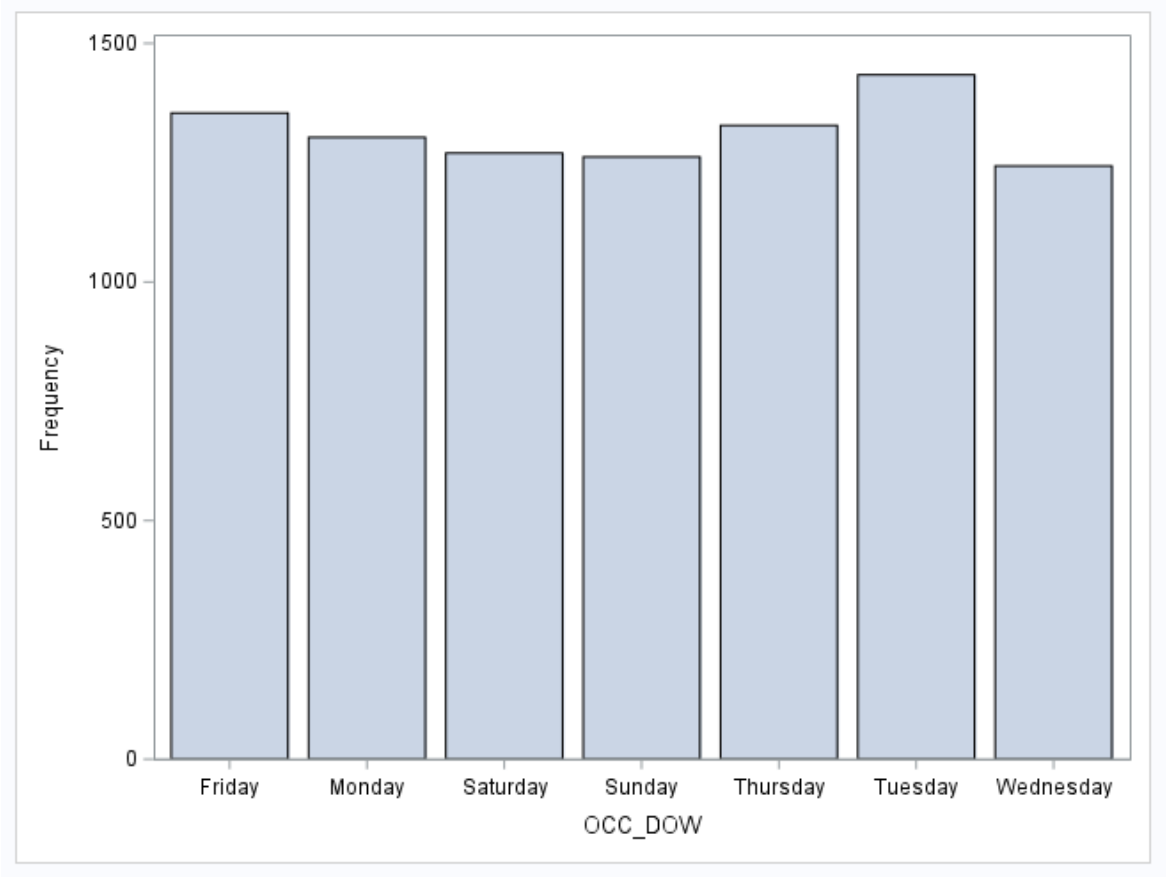
Univariate - Qualitative

Variable:OCC_DOW (Day of week the offence occurred)

*Q4 - On which day of the week are bicycles most frequently theft?

The FREQ Procedure

OCC_DOW	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Friday	1354	14.73	1354	14.73
Monday	1303	14.17	2657	28.90
Saturday	1270	13.81	3927	42.71
Sunday	1262	13.73	5189	56.44
Thursday	1328	14.44	6517	70.88
Tuesday	1434	15.60	7951	86.48
Wednesday	1243	13.52	9194	100.00



Conclusion:

In this analysis of the **OCC_DOW** variable, we observe that the difference between days of week is almost the same percentage between 13% and 16%.

Q5 - Bicycle Theft Type

We analyzed the number of bicycle thefts for each type of bike using the PROC FREQ statement. This allowed us to calculate the frequency and percentage of thefts for each type.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: BIKE_TYPE

Visualization Method: Vertical Bar Chart (VBAR)

Analysis Technique: PROC FREQ

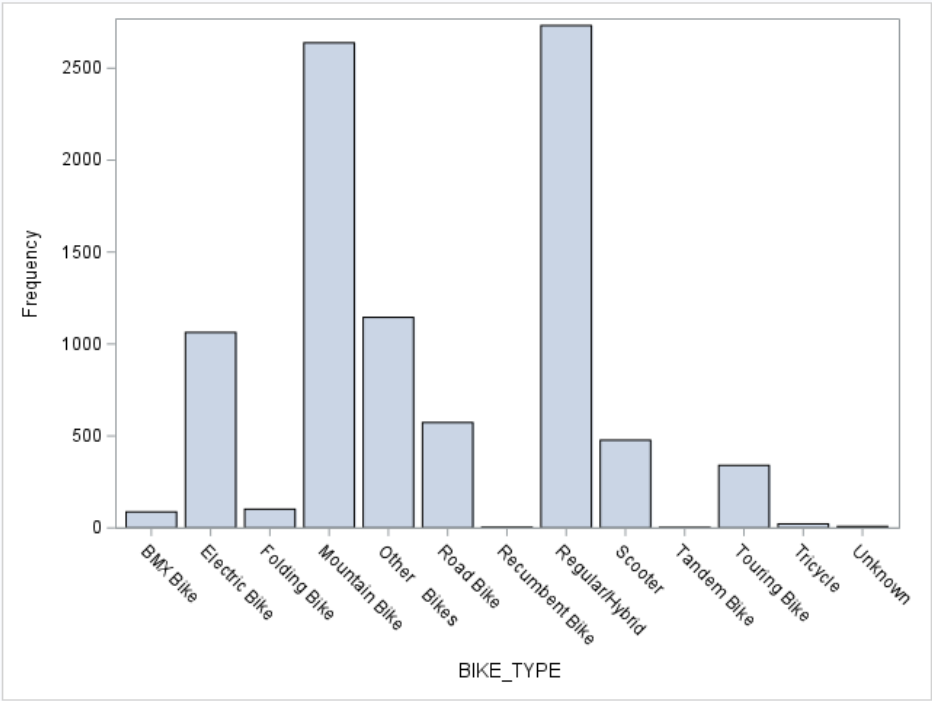
Univariate - Qualitative

Variable:BIKE_TYPE (Type of Bicycle)

***Q5 - Whats type of bike is usually is more theft?**

The FREQ Procedure

BIKE_TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
BMX Bike	87	0.95	87	0.95
Electric Bike	1063	11.56	1150	12.51
Folding Bike	102	1.11	1252	13.62
Mountain Bike	2638	28.69	3890	42.31
Other Bikes	1145	12.45	5035	54.76
Road Bike	573	6.23	5608	61.00
Recumbent Bike	4	0.04	5612	61.04
Regular/Hybrid	2732	29.72	8344	90.75
Scooter	477	5.19	8821	95.94
Tandem Bike	3	0.03	8824	95.98
Touring Bike	340	3.70	9164	99.67
Tricycle	22	0.24	9186	99.91
Unknown	8	0.09	9194	100.00



Conclusion:

By analyzing the frequency table, we can see that Regular/Hybrid bikes and Mountain bikes account for a total of 58.41% of bike thefts.

Q6 - Bicycle Theft by Premise Type

To determine the most common premises where bicycle thefts occur, we used the PROC FREQ statement. This allowed us to calculate the frequency and percentage of theft incidents by premise type.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: PREMISE_TYPE

Visualization Method: Vertical Bar Chart (VBAR)

Analysis Technique: PROC FREQ

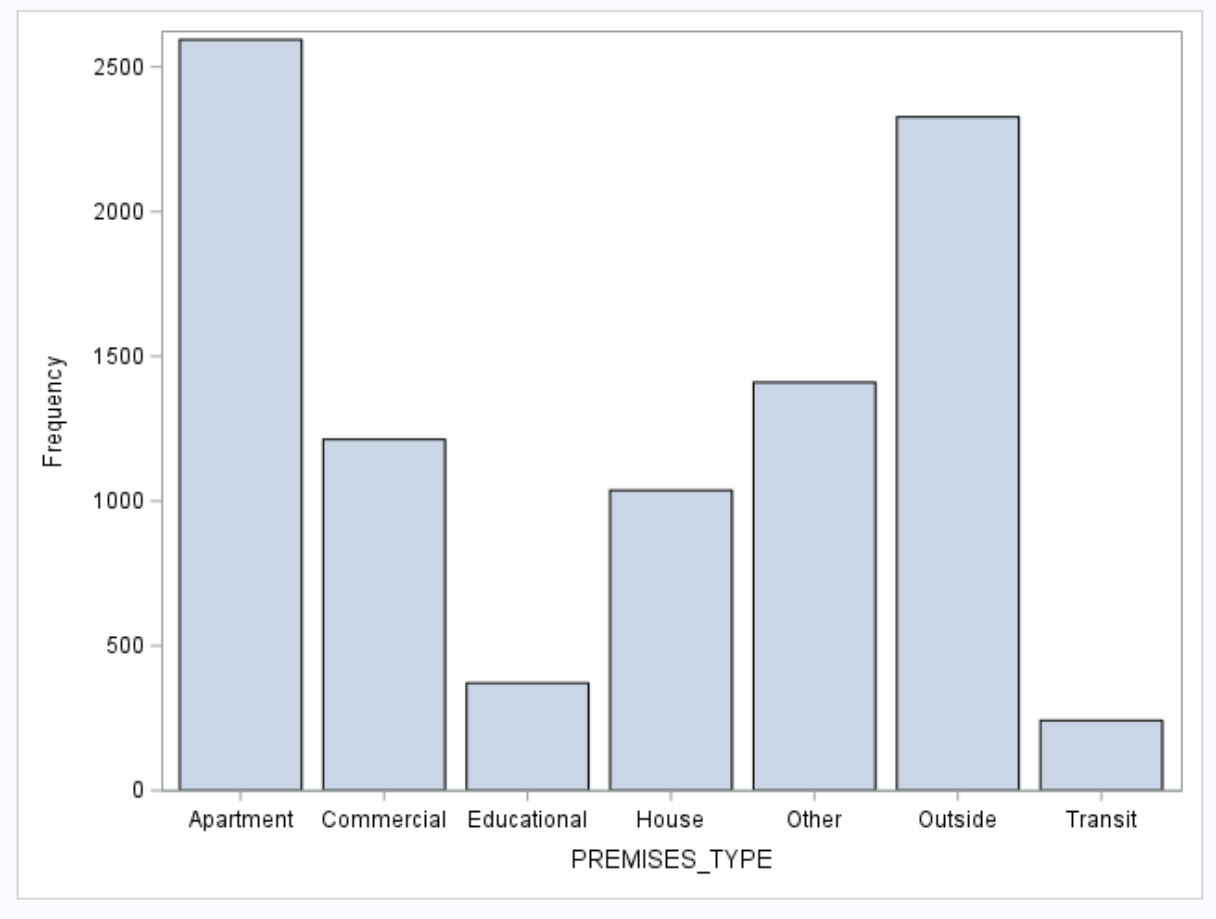
Univariate - Qualitative

Variable:PREMISES_TYPE (Premises Type of Offence)

Q6 - What types of premises are bicycles most frequently stolen from?

The FREQ Procedure

PREMISES_TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Apartment	3161	27.82	3161	27.82
Commercial	1455	12.80	4616	40.62
Educational	415	3.65	5031	44.28
House	1317	11.59	6348	55.87
Other	1741	15.32	8089	71.19
Outside	2986	26.28	11075	97.47
Transit	288	2.53	11363	100.00



Conclusion:

By analyzing the frequency table, we can see that 54.10% of bike thefts is from Apartment and Outside places.

Q7 - What is the average price of a stolen bicycle?

To answer this question, we will analyze the BIKE_COST variable, which represents the cost of stolen bikes.

Analysis Overview:

Type: Univariate Analysis

Numerical Variable: BIKE_COST (Continuous)

Visualization: Box Plot (Horizontal Box Plot)

Analysis Method: Univariate

Univariate - Quantitative

Variable:BIKE_COST (Cost of Bicycle)

Q7 - What is the average price of a theft bicycle?

The UNIVARIATE Procedure
Variable: BIKE_COST

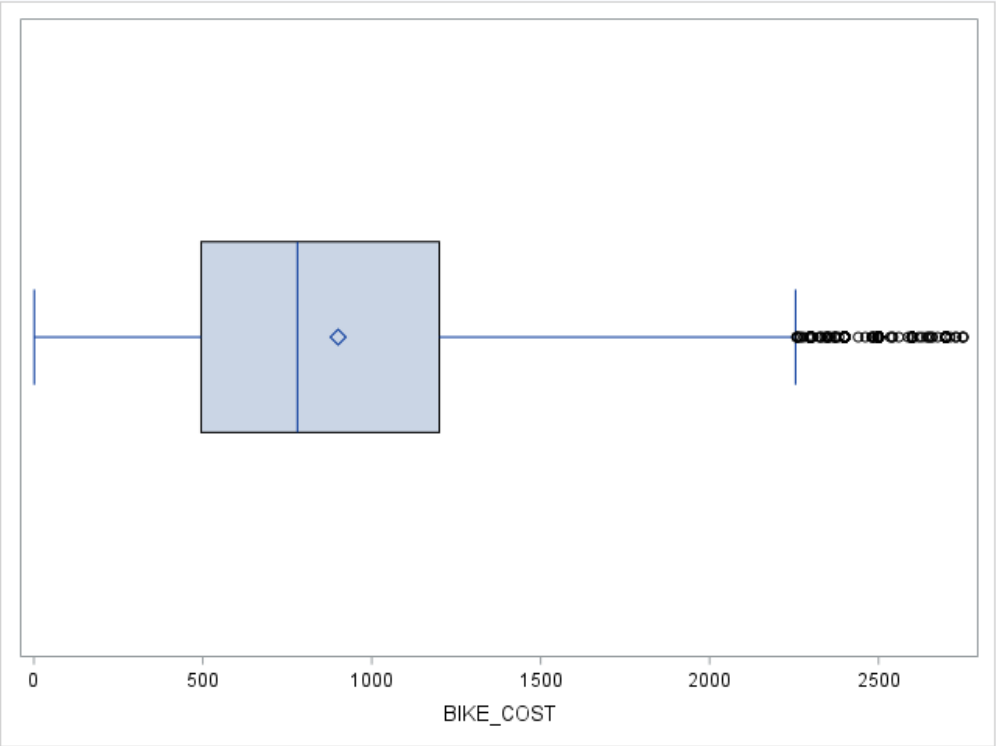
Moments			
N	9194	Sum Weights	9194
Mean	900.533663	Sum Observations	8279506.5
Std Deviation	604.312406	Variance	365193.484
Skewness	1.01963653	Kurtosis	0.48767011
Uncorrected SS	1.08132E10	Corrected SS	3357223697
Coeff Variation	67.1060317	Std Error Mean	6.30244784

Basic Statistical Measures			
Location		Variability	
Mean	900.534	Std Deviation	604.31241
Median	780.000	Variance	365193
Mode	1000.000	Range	2749
		Interquartile Range	705.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	142.8863	Pr > |t|	<.0001
Sign	M	4597	Pr >= |M|	<.0001
Signed Rank	S	21134708	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2750
99%	2500
95%	2200
90%	1900
75% Q3	1200
50% Median	780
25% Q1	495
10%	250
5%	168
1%	100
0% Min	1

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	4264	2728	4398
3	7186	2730	6542
5	3983	2750	6197
12	5863	2750	6655
20	4916	2750	8263



Conclusion:

For the **BIKE_COST** variable, I performed a univariate analysis in SAS, which allowed us to examine specific values. Based on this analysis, we can conclude that the average cost of a stolen bike is $900.53. The five-number summary reveals that the median cost is $980. Additionally, 25% (Q1) of our observations have a cost of $495, while 75% (Q3) have a cost of $1,200. The minimum reported price was $1 CAD, and the maximum was $2,750.

Q8 - What is the average maximum speed of stolen bicycles?

To answer this question, we will analyze the **BIKE_SPEED** variable, which represents the minimum, maximum and average speed of the stolen bicycles.

Analysis Overview:

Type: Univariate Analysis

Numerical Variable: BIKE_SPEED (Continuous)

Visualization: Box Plot (Horizontal Box Plot)

Analysis Method: Means

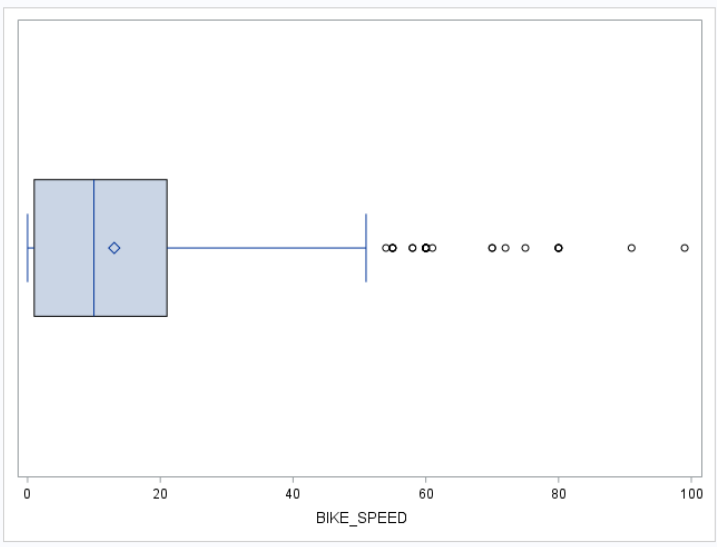
Univariate - Quantitative

Variable:BIKE_SPEED (Speed max of Bicycle)

***Q8 - What's the avarage of speed max of bycicles theft?**

The MEANS Procedure

Analysis Variable : BIKE_SPEED				
N	Mean	Std Dev	Minimum	Maximum
9194	13.0602567	11.2040818	0	99.0000000



Conclusion:

Based on means table, we can see that the average of speed of bike stolen is 13.06 km/h max.

Q9 - How many bicycles were recovered?

To answer this question, we need to use the **PROC FREQ** statement to analyze the frequency and calculate the percentage of bicycles that were stolen by year.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: STATUS

Visualization Method: Pie Chart / Vertical Bar Chart (VBAR)

Analysis Method: Frequency Distribution (PROC FREQ)

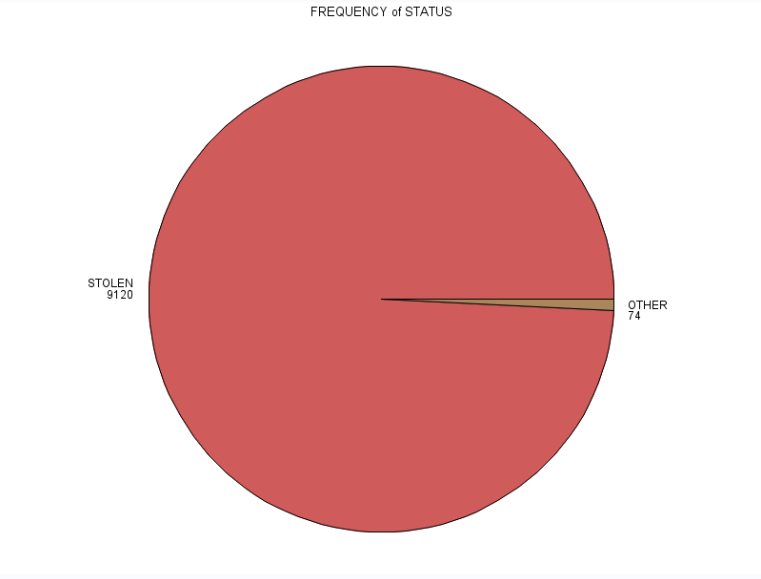
Univariate - Qualitative

Variable:STATUS (Status of Bicycle)

Q9 - How many bicyles was recovered ?

The FREQ Procedure

STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
RECOVERED	43	0.47	43	0.47
STOLEN	9120	99.20	9163	99.66
UNKNOWN	31	0.34	9194	100.00



Conclusion:

Only 0.47% of bikes thefts was recovered.

Q10 - How many days, on average, pass between the date a bicycle is stolen and the date the theft is reported?

To answer this question, we will analyze the **NDAYS** variable, which represents the number of days between the theft and the report date.

Analysis Overview:

Type: Univariate - Qualitative

Categorical Variable: NDAYS

Visualization: Bar Chart (VBAR)

Analysis Method: Frequency Distribution (PROC FREQ)

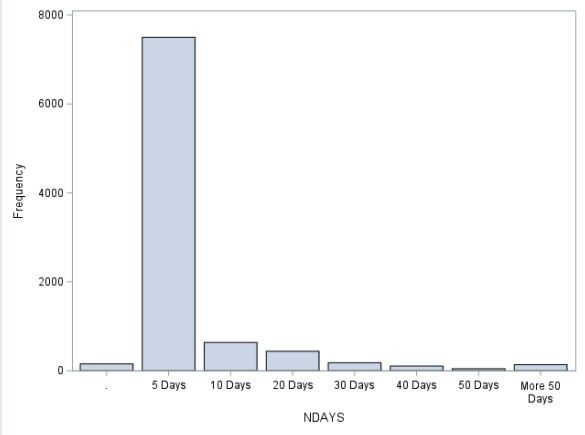
Univariate - Qualitative

Variable:STATUS (Status of Bicycle)

Q10 - How many days, on average, pass between the date a bicycle is stolen and the date the theft is reported?

The FREQ Procedure

NDAYS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	154	1.68	154	1.68
5 Days	7499	81.56	7653	83.24
10 Days	636	6.92	8289	90.16
20 Days	438	4.76	8727	94.92
30 Days	180	1.96	8907	96.88
40 Days	105	1.14	9012	98.02
50 Days	45	0.49	9057	98.51
More 50 Days	137	1.49	9194	100.00



Conclusion:

Based on results from frequencies table we can say that 81.56% reported the theft in 5 days maximum after the theft, and 1.49% reported just after 50 days.

Bivariate Analysis

Q1 - What has been the growth in e-bike thefts from 2020 to 2024?

To answer this question, we will examine the relationship between e-bike thefts and the years reported, focusing on the **BIKE_TYPE** and **REPORT_YEAR** variables.

Analysis Overview:

Type: Bivariate

Categorical vs. Categorical Variables: BIKE_TYPE * REPORT_YEAR

Visualization: Bar Chart (VBAR)

Test of Independence: Contingency Table

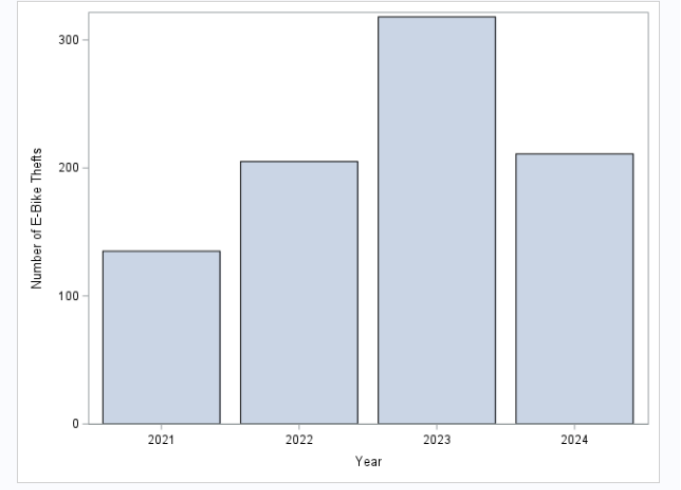
Bivariate - CONTIGENCE TABLE

Variables: BIKE_TYPE * REPORT_YEAR

Q1 - What has been the growth in e-bike thefts from 2020 to 2024/09?

The FREQ Procedure

Frequency Percent	Table of BIKE_TYPE by REPORT_YEAR				
	REPORT_YEAR				
BIKE_TYPE	2021	2022	2023	2024	Total
EL	135 15.54	205 23.59	318 36.59	211 24.28	869 100.00
Total	135 15.54	205 23.59	318 36.59	211 24.28	869 100.00



Conclusion:

E-bike thefts have shown substantial growth from **2021 to 2023** with an increase of **135.56%**.

However, there has been a decline of **33.65%** in 2024 (up to September) compared to 2023.

Q2 - Does the day of the week affect the cost of the bike that was stolen?

To answer this question, we will analyze the relationship between the day of the week and the cost of the stolen bike.

Analysis Overview:

Type: Bivariate

Categorical Variables: OCC_DOW (day of the week) and BIKE_COST (after formatting)

Visualization: Clustered Bar Chart

Test of Independence: Contingency Table

Bivariate - CONTIGENCE TABLE

Variables: OCC_DOW * BIKE_COST

Q2 -The day of the week is affected the cost of bike was theft ?

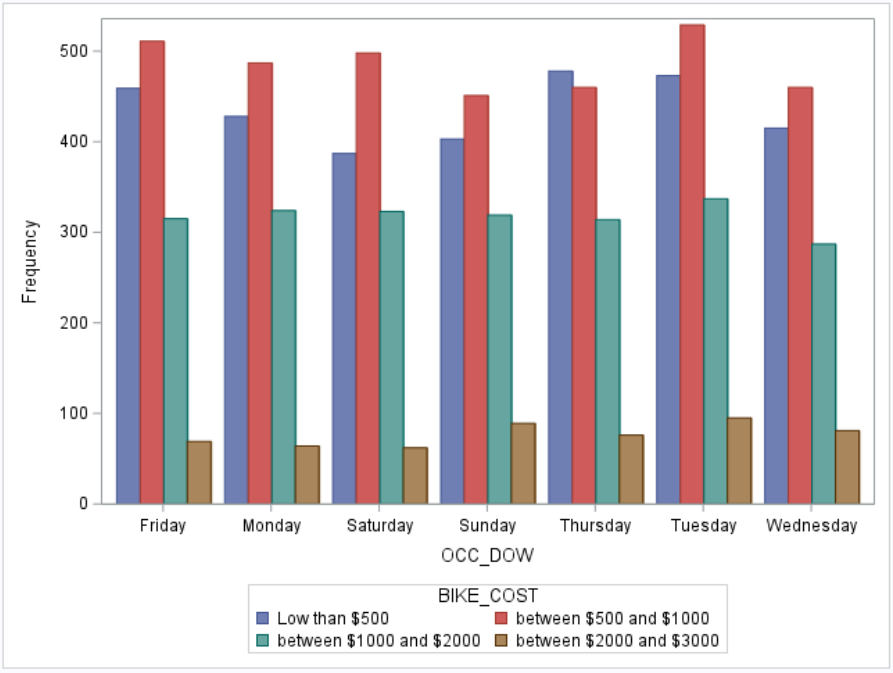
The FREQ Procedure

Frequency Percent	Table of OCC_DOW by BIKE_COST				
	BIKE_COST				
OCC_DOW	Low than $500	between $500 and $1000	between $1000 and $2000	between $2000 and $3000	Total
Friday	459 4.99	511 5.56	315 3.43	69 0.75	1354 14.73
Monday	428 4.66	487 5.30	324 3.52	64 0.70	1303 14.17
Saturday	387 4.21	498 5.42	323 3.51	62 0.67	1270 13.81
Sunday	403 4.38	451 4.91	319 3.47	89 0.97	1262 13.73
Thursday	478 5.20	460 5.00	314 3.42	76 0.83	1328 14.44
Tuesday	473 5.14	529 5.75	337 3.67	95 1.03	1434 15.60
Wednesday	415 4.51	460 5.00	287 3.12	81 0.88	1243 13.52
Total	3043 33.10	3396 36.94	2219 24.14	536 5.83	9194 100.00

Statistics for Table of OCC_DOW by BIKE_COST

Statistic	DF	Value	Prob
Chi-Square	18	25.4603	0.1128
Likelihood Ratio Chi-Square	18	25.4104	0.1140
Mantel-Haenszel Chi-Square	1	1.2511	0.2634
Phi Coefficient		0.0526	
Contingency Coefficient		0.0526	
Cramer's V		0.0304	

Sample Size = 9194



Conclusion:

Across all days of the week, most thefts occur in the $500–$1,000 cost range, accounting for 36.94% of total thefts.

The $2,000–$3,000 range consistently has the fewest thefts (5.83% total).

Tuesday has the highest overall percentage of thefts (15.60%), while Wednesday has the lowest (13.52%).

The patterns are generally consistent across all days, with slight variations in percentages

Since the p-values are greater than α=0.05\alpha = 0.05α=0.05, we **fail to reject the null hypothesis**. This suggests that there is no statistically significant association between the day of the week (OCC_DOW) and the cost of the stolen bike (BIKE_COST).

Q3 - Does the time of day affect the cost of the bike that was stolen?

To answer this question, we will analyze the relationship between the time of day and the cost of the stolen bike.

Analysis Overview:

Type: Bivariate

Categorical Variables: BIKE_COST (after formatting) and OCC_HOUR (after formatting)

Visualization: Clustered Bar Chart

Test of Independence: Chi-Square Test

Bivariate - CHI SQUARE

Variables: OCC_HOUR * BIKE_COST

Q3 - Is the period of day affect the bike cost was theft ?

The FREQ Procedure

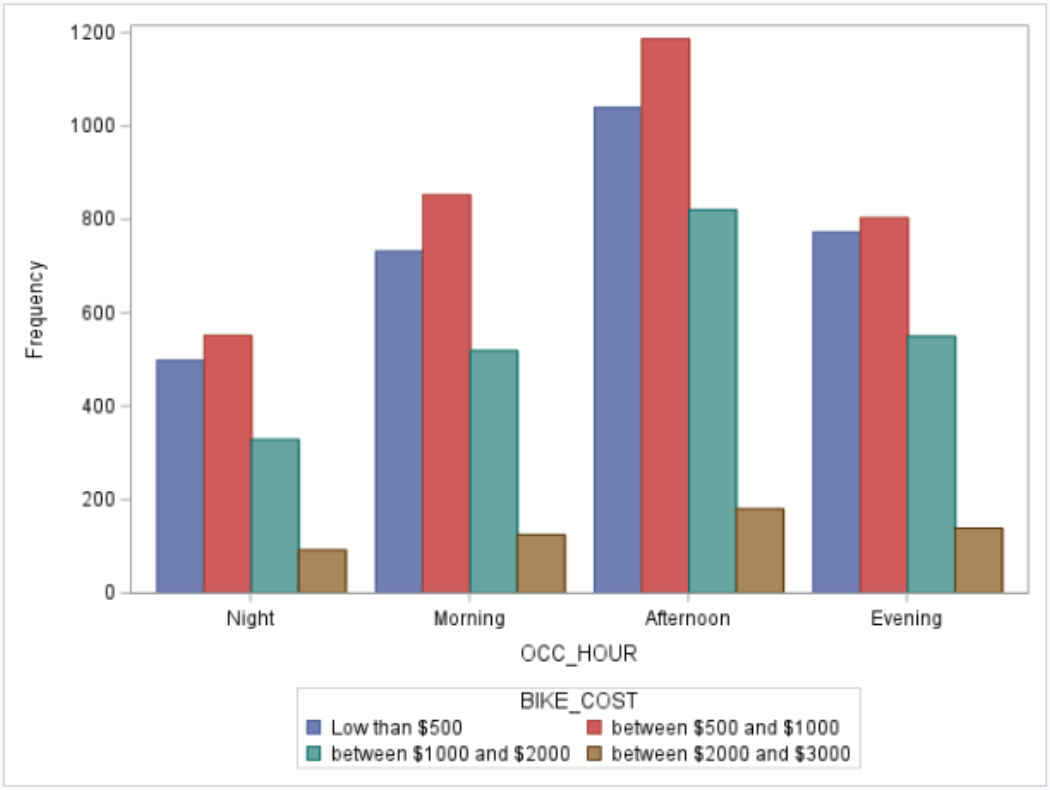
Table of OCC_HOUR by BIKE_COST

Frequency Percent	BIKE_COST				
OCC_HOUR	Low than $500	between $500 and $1000	between $1000 and $2000	between $2000 and $3000	Total
Night	498 5.42	552 6.00	329 3.58	92 1.00	1471 16.00
Morning	732 7.96	853 9.28	519 5.64	125 1.36	2229 24.24
Afternoon	1040 11.31	1187 12.91	821 8.93	180 1.96	3228 35.11
Evening	773 8.41	804 8.74	550 5.98	139 1.51	2266 24.65
Total	3043 33.10	3396 36.94	2219 24.14	536 5.83	9194 100.00

Statistics for Table of OCC_HOUR by BIKE_COST

Statistic	DF	Value	Prob
Chi-Square	9	10.5004	0.3115
Likelihood Ratio Chi-Square	9	10.5104	0.3108
Mantel-Haenszel Chi-Square	1	0.4144	0.5197
Phi Coefficient		0.0338	
Contingency Coefficient		0.0338	
Cramer's V		0.0195	

Sample Size = 9194



Conclusion:

Since the p-value (0.3115) is greater than the significance level α=0.05, **we fail to reject the null hypothesis**. Therefore, we conclude that there is no significant association between the period of the day and the cost of the bike.

Q4 - Is there any relationship between the premises where the bike was stolen and the season of the year?

To answer this question, we will analyze the relationship between the type of premises and the month of the theft.

Analysis Overview:

Type: Bivariate

Categorical Variables: PREMISES_TYPE and OCC_MONTH

Visualization: Bar Chart (VBAR)

Test of Independence: Chi-Square Test

Bivariate - CHI SQUARE

Variables:PREMISES_TYPE*OCC_MONTH

Q4 - Is any relatioshinp between premises where bike wa theft and season of year?

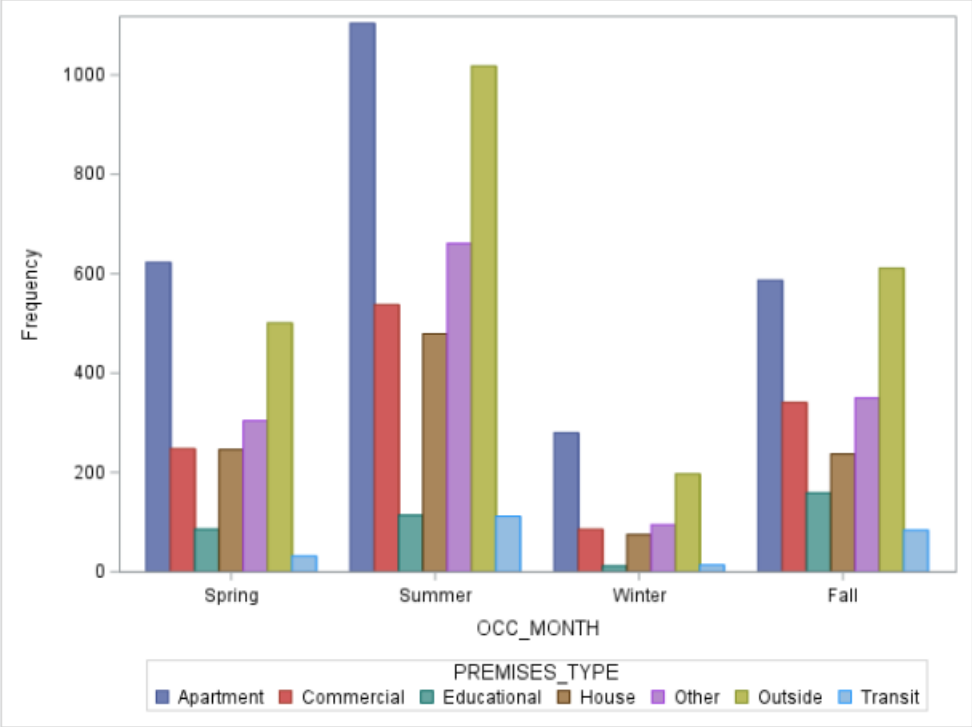
The FREQ Procedure

Frequency Percent	Table of PREMISES_TYPE by OCC_MONTH				
	OCC_MONTH				
PREMISES_TYPE	Spring	Summer	Winter	Fall	Total
Apartment	623 6.78	1104 12.01	280 3.05	587 6.38	2594 28.21
Commercial	248 2.70	538 5.85	86 0.94	341 3.71	1213 13.19
Educational	86 0.94	114 1.24	12 0.13	159 1.73	371 4.04
House	246 2.68	479 5.21	75 0.82	237 2.58	1037 11.28
Other	304 3.31	661 7.19	95 1.03	350 3.81	1410 15.34
Outside	501 5.45	1018 11.07	197 2.14	611 6.65	2327 25.31
Transit	32 0.35	112 1.22	14 0.15	84 0.91	242 2.63
Total	2040 22.19	4026 43.79	759 8.26	2369 25.77	9194 100.00

Statistics for Table of PREMISES_TYPE by OCC_MONTH

Statistic	DF	Value	Prob
Chi-Square	18	143.4556	<.0001
Likelihood Ratio Chi-Square	18	141.4311	<.0001
Mantel-Haenszel Chi-Square	1	3.7859	0.0517
Phi Coefficient		0.1249	
Contingency Coefficient		0.1239	
Cramer's V		0.0721	

Sample Size = 9194



Conclusion:

Since the p-value is less than significance level α=0.05, **we can reject the null hypothesis**, and conclude that there is an association between premises type and season of year.

Q5 - Is there any correlation between the maximum speed of the bike and its cost?

To answer this question, we will analyze the relationship between the bike's cost and its maximum speed.

Analysis Overview:

Type: Bivariate

Continuous Variables: BIKE_COST and BIKE_SPEED

Visualization: Scatter Plot Matrix

Test of Independence: Correlation (PROC CORR)

Bivariate - CORRELATION

Variable:BIKE_SPEED VS BIKE_COST

Q5 - Is any correlation between speed max of bike and cost?

The CORR Procedure

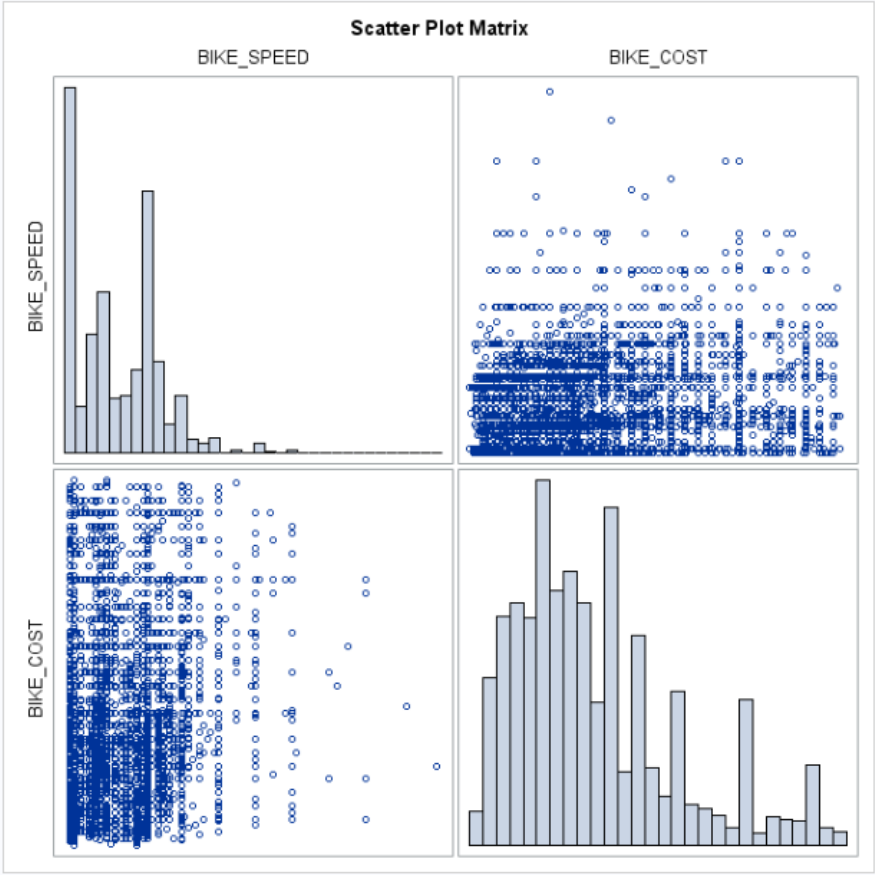
2 Variables:	BIKE_SPEED BIKE_COST

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
BIKE_SPEED	9194	13.06026	11.20408	120076	0	99.00000
BIKE_COST	9194	900.53366	604.31241	8279506	1.00000	2750

Pearson Correlation Coefficients, N = 9194
Prob > |r| under H0: Rho=0

	BIKE_SPEED	BIKE_COST
BIKE_SPEED	1.00000	0.04873 <.0001
BIKE_COST	0.04873 <.0001	1.00000



Conclusion:

Correlation Value: The correlation coefficient is 0.04873, indicating a very weak linear relationship between bike speed and bike cost. This suggests that bike speed and cost are almost independent, with no meaningful relationship.

P-value (0.001): Since the p-value is slightly below 0.05, the result is statistically significant at the 5% level. This means that while the correlation is very close to zero, there is weak evidence of a small inverse relationship.

Q6 - Is there any relationship between the average bike cost and weekdays versus weekends?

To answer this question, we will analyze the relationship between the average bike cost and the day of the week.

Analysis Overview:

Type: Bivariate

Continuous Variable: BIKE_COST (Dependent)

Categorical Variable: OCC_DOW (Independent)

Test of Independence: T-Test

Bivariate - T-TEST

Variable:BIKE_COST VS OCC_DOW

Q6 - Any relationship between the avarage of bike cost and week days and weekend?

The TTEST Procedure

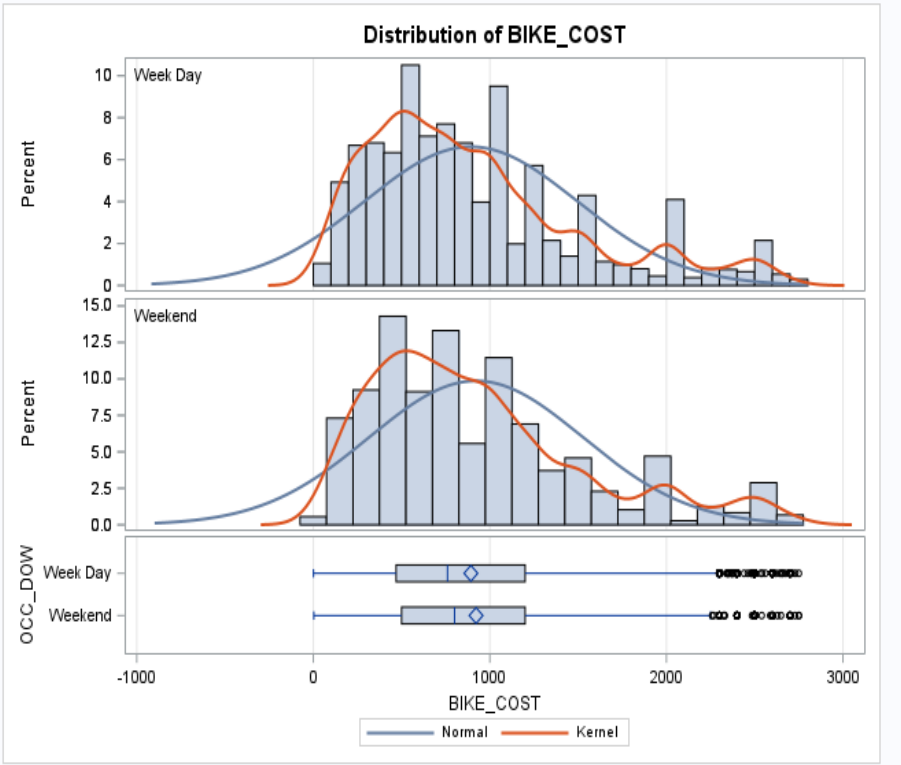
Variable: BIKE_COST

OCC_DOW	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Week Day		6662	892.5	603.2	7.3903	1.0000	2750.0
Weekend		2532	921.8	606.8	12.0595	3.0000	2750.0
Diff (1-2)	Pooled		-29.2870	604.2	14.1059		
Diff (1-2)	Satterthwaite		-29.2870		14.1438		

OCC_DOW	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Week Day		892.5	878.0	907.0	603.2	593.1	613.6
Weekend		921.8	898.1	945.4	606.8	590.6	624.0
Diff (1-2)	Pooled	-29.2870	-56.9377	-1.6363	604.2	595.6	613.1
Diff (1-2)	Satterthwaite	-29.2870	-57.0158	-1.5582			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	9192	-2.08	0.0379
Satterthwaite	Unequal	4545.4	-2.07	0.0384

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2531	6661	1.01	0.7134



Histogram Analysis

Both distributions are slightly right-skewed, meaning higher-cost bikes are less frequent but present in both categories.
The kernel density plot shows the general trend of the data, while the normal curve (blue line) suggests the data is not perfectly normal.

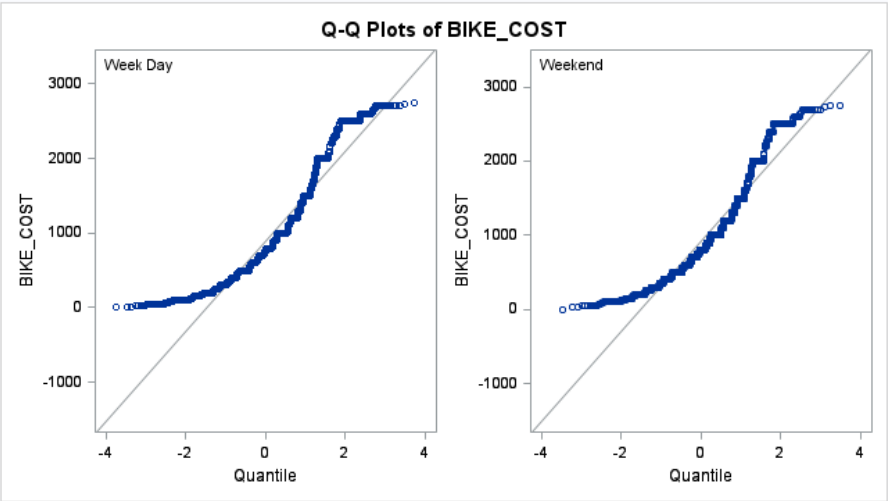
Comparison of Means

Weekdays seem to have a slightly higher concentration of lower-cost bikes, with the mid-range

Weekends show a broader spread, indicating that more high-cost bikes involved in thefts compared to weekdays.

Box Plot Analysis

The medians appear slightly higher for weekends compared to weekdays, also the interquartile (IQR) is wider on weekends, suggesting higher variability in bike cost during theses days. Both categories show outliers, but in weekends having more extreme values.



Q-Q Plots

The values for both categoric isn't normally distributed in the tails, in the middle of the range are consistent with normal distribution.

Conclusion:

The difference in means is 29.2870 – weekends having a slightly higher average bike cost.

Base that we now the both standard deviation is similar we should look at Pooled confidence interval.

P-value: is less the 0.05 this indicate a statistically significant difference in the average bike cost between weekdays and weekend.

"Formal" hypothesis, Folded F – test has p-value that indicating that the variances between the two groups are not significantly different. Also because p-value is grater then (0.5) we fail to eject the null hypothesis.

Q7 – Can we say that the price of the bike changes if the speed changes as well?

Analysis Overview:

Type: Bivariate

Continuous Variable: BIKE_COST (Dependent)

Continuous Variable: BIKE_SPEED (Independent)

Test of Independence: ANOVA

Bivariate - ANOVA

Variable:BIKE_COST = BIKE_SPEED

Can we say that the price of the bike changes if the speed changes as well?

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
BIKE_SPEED	4	100 k/h max 25 k/h max 50 k/h max No motorized

Number of Observations Read	9194
Number of Observations Used	9194

Bivariate - ANOVA

Variable:BIKE_COST = BIKE_SPEED

Can we say that the price of the bike changes if the speed changes as well?

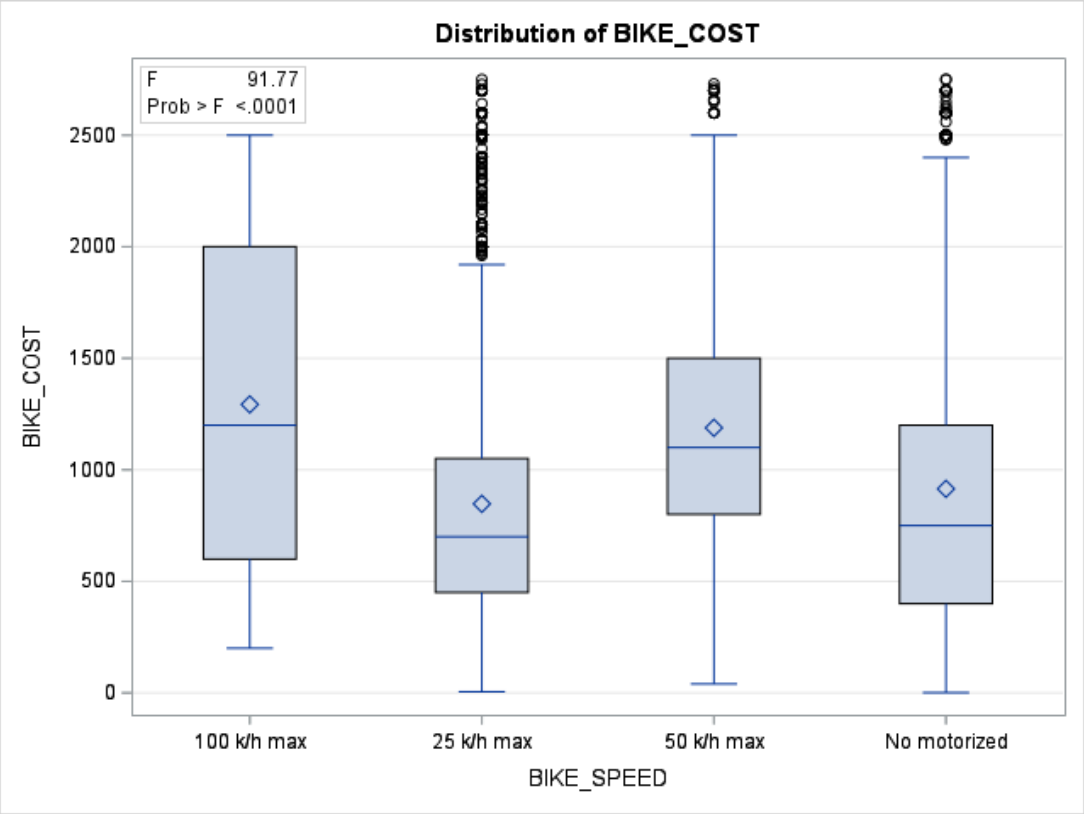
The ANOVA Procedure

Dependent Variable: BIKE_COST

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	97648988	32549663	91.77	<.0001
Error	9190	3259574708	354687		
Corrected Total	9193	3357223697			

R-Square	Coeff Var	Root MSE	BIKE_COST Mean
0.029086	66.13369	595.5562	900.5337

Source	DF	Anova SS	Mean Square	F Value	Pr > F
BIKE_SPEED	3	97648988.43	32549662.81	91.77	<.0001



Conclusion:

R-Square: 0.0291

This indicates that 0.0291 of the variance in BIKE_COST is explained by BIKE_SPEED. This suggests that while the relationship is statistically significant, bike speed only explains a small portion of the variability in bike cost.

Based on the p-value (<.0001), you can conclude that there is a statistically significant relationship between bike cost and bike speed.

Conclusion

After my analysis, I can conclude that for all univariate analyses, I was able to draw meaningful conclusions about certain variables, such as which season of the year, time of day, or day of the week bicycles were most often stolen. Based on these conclusions, we can make predictions. On the other hand, for the bivariate analysis, the data wasn't properly cleaned, which compromised my univariate analysis. Since this dataset does not follow a normal distribution, we cannot rely on the results from this type of analysis to make accurate predictions.