Bicycle Store data analysis project

Prepared by David Schied

As an avid bicyclist, I thought it would be fun to practice my data analysis skills using a bicycle store dataset I found on Kaggle. The link to the dataset page is:

https://www.kaggle.com/datasets/nikhilchandra78/bicyclestoredata

The dataset does not describe any real problem that is trying to be solved. However, it gave me the opportunity to apply the data sensemaking skills I've learned by reading several of Stephen Few's books on the subject (Show Me the Numbers & Now You See It). Instead of solving a problem, I've suggested questions that might be explored if other data could be collected.

My Method

I used Python Pandas to perform most of the data gathering, as well as Matplotlib for graphing as much as possible. Python code was run on a Jupyter Notebook on Kaggle.

Excel was used to provide better presentation of tables, and for graphing where I could not get what I wanted from Python.

General Data Overview

Data Sample (Pandas Dataframe):

	Date	Day	Month	Year	Customer_Age	Age_Group	Customer_Gender	Country	State	Product_Category	Sub_Category	Product	Order_Quantity	Unit_Cost	Unit_Price	Profit	Cost	Revenue
0	2013-11-26	26	November	2013	19	Youth (<25)	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack - 4-Bike	8	45	120	590	360	950
1	2015-11-26	26	November	2015	19	Youth (<25)	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack - 4-Bike	8	45	120	590	360	950
2	2014-03-23	23	March	2014	49	Adults (35-64)	M	Australia	New South Wales	Accessories	Bike Racks	Hitch Rack - 4-Bike	23	45	120	1366	1035	2401
3	2016-03-23	23	March	2016	49	Adults (35-64)	M	Australia	New South Wales	Accessories	Bike Racks	Hitch Rack - 4-Bike	20	45	120	1188	900	2088
4	2014-05-15	15	May	2014	47	Adults (35-64)	F	Australia	New South Wales	Accessories	Bike Racks	Hitch Rack - 4-Bike	4	45	120	238	180	418

Dataframe Information.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113036 entries, 0 to 113035
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Date              113036 non-null  object
 1   Day               113036 non-null  int64
 2   Month             113036 non-null  object
 3   Year              113036 non-null  int64
 4   Customer_Age      113036 non-null  int64
 5   Age_Group         113036 non-null  object
 6   Customer_Gender   113036 non-null  object
 7   Country           113036 non-null  object
 8   State             113036 non-null  object
 9   Product_Category  113036 non-null  object
 10  Sub_Category      113036 non-null  object
 11  Product           113036 non-null  object
 12  Order_Quantity    113036 non-null  int64
 13  Unit_Cost         113036 non-null  int64
 14  Unit_Price        113036 non-null  int64
 15  Profit            113036 non-null  int64
 16  Cost              113036 non-null  int64
 17  Revenue           113036 non-null  int64
dtypes: int64(9), object(9)
```

NOTE: Date is an Object type and will need to be converted to Datetime format. There are no missing values.

Descriptive Statistics on select data columns.

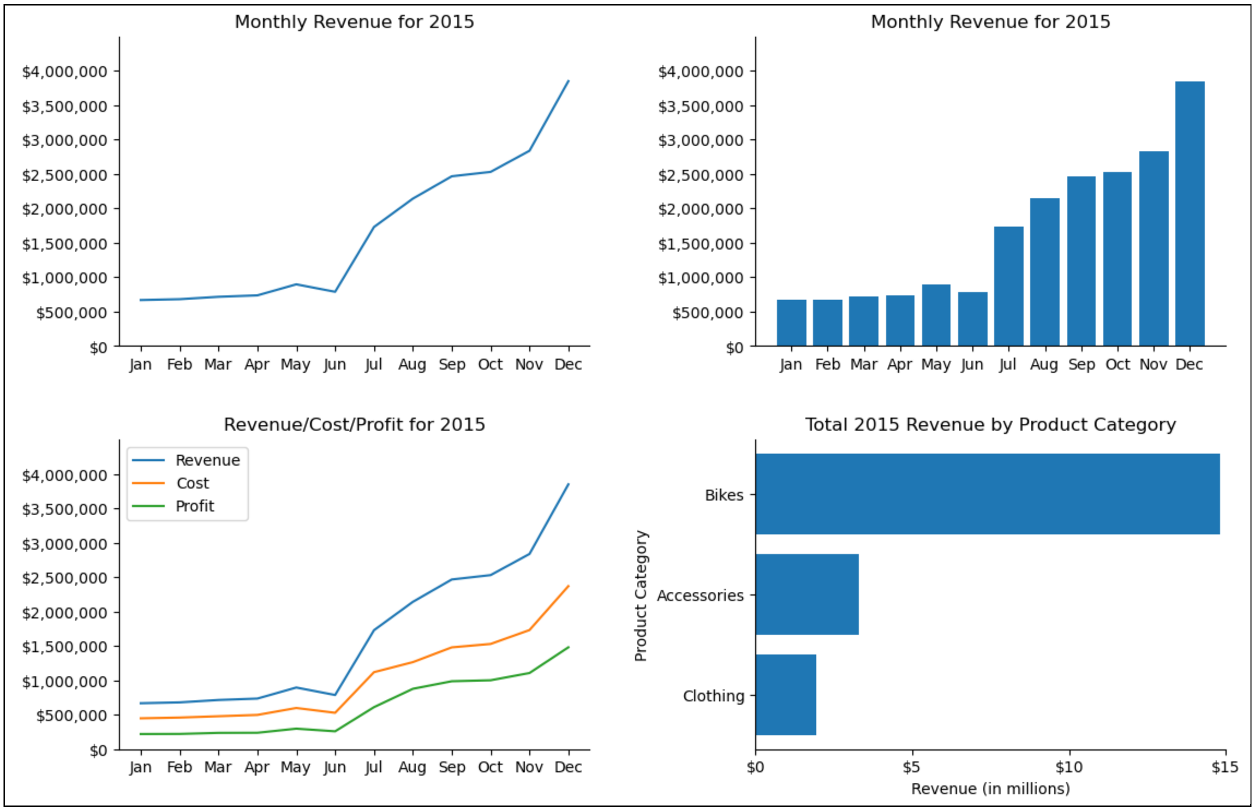
	Order_Quantity	Unit_Cost	Unit_Price	Profit	Cost	Revenue
count	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000
mean	11.901660	267.296366	452.938427	285.051665	469.318695	754.370360
std	9.561857	549.835483	922.071219	453.887443	884.866118	1309.094674
min	1.000000	1.000000	2.000000	-30.000000	1.000000	2.000000
25%	2.000000	2.000000	5.000000	29.000000	28.000000	63.000000
50%	10.000000	9.000000	24.000000	101.000000	108.000000	223.000000
75%	20.000000	42.000000	70.000000	358.000000	432.000000	800.000000
max	32.000000	2171.000000	3578.000000	15096.000000	42978.000000	58074.000000

Beginning and Ending Dates of the Dataset

```
Beginning date 2011-01-01 00:00:00
Ending date 2016-07-31 00:00:00
```

<u>**Views and Analysis of the Data**</u>

I arbitrarily selected 2015 as a full year to calculate some basic high-level views.



Graphs were created as subplots in Python Matplotlib.

Questions to explore further:

- How does revenue, cost, and profit compare to prior years?
- How should we further slice the data? Products, customer age groups, countries?

Let’s look at the top performing and underperforming products:

Top 10 Profitable Products for 2015	Profit	Profit Contribution
Mountain-200 Black, 38	$360,619	
Mountain-200 Silver, 42	$335,263	
Road-150 Red, 62	$324,152	
Mountain-200 Silver, 38	$297,130	
Mountain-200 Black, 42	$284,063	
Road-150 Red, 56	$272,863	
Sport-100 Helmet, Red	$259,914	
Sport-100 Helmet, Black	$256,857	
Road-150 Red, 52	$256,791	
Mountain-200 Black, 46	$236,212	
Total profit from Top 10 Products	$2,883,864	38.3%

Bottom 10 Profitable Products for 2015	Profit	Profit Contribution
Mountain-500 Black, 42	$4,055	
Mountain-500 Silver, 52	$4,006	
Road-650 Black, 48	$3,984	
Touring-3000 Yellow, 58	$3,760	
Touring-3000 Blue, 50	$3,005	
Road-650 Black, 44	$2,802	
Mountain-500 Black, 52	$2,048	
Mountain-500 Silver, 48	$2,031	
Road-650 Red, 44	$1,275	
Road-650 Red, 52	$637	
Total profit from Bottom 10 Products	$27,603	0.4%

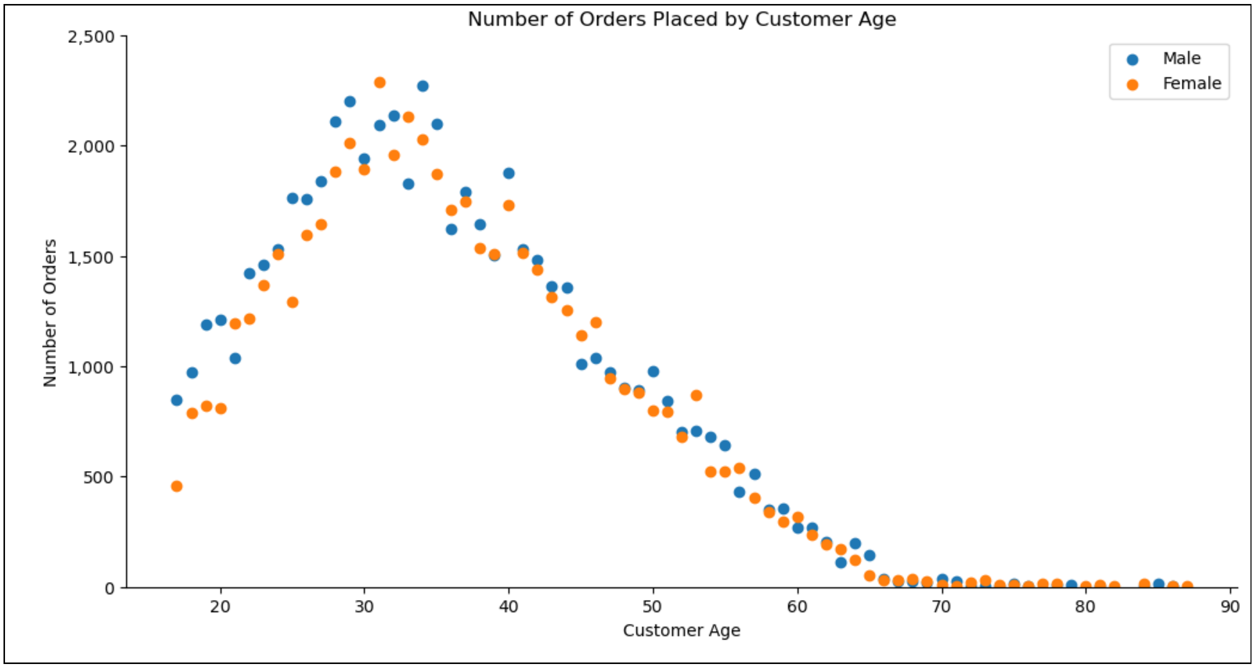
Total profit for 2015 $7,528,563

Data summarized using Panda, exported to Excel for formatting.

Questions to explore further:

- What should we do about the underperforming products?
- Can we sell more of the high performing products?

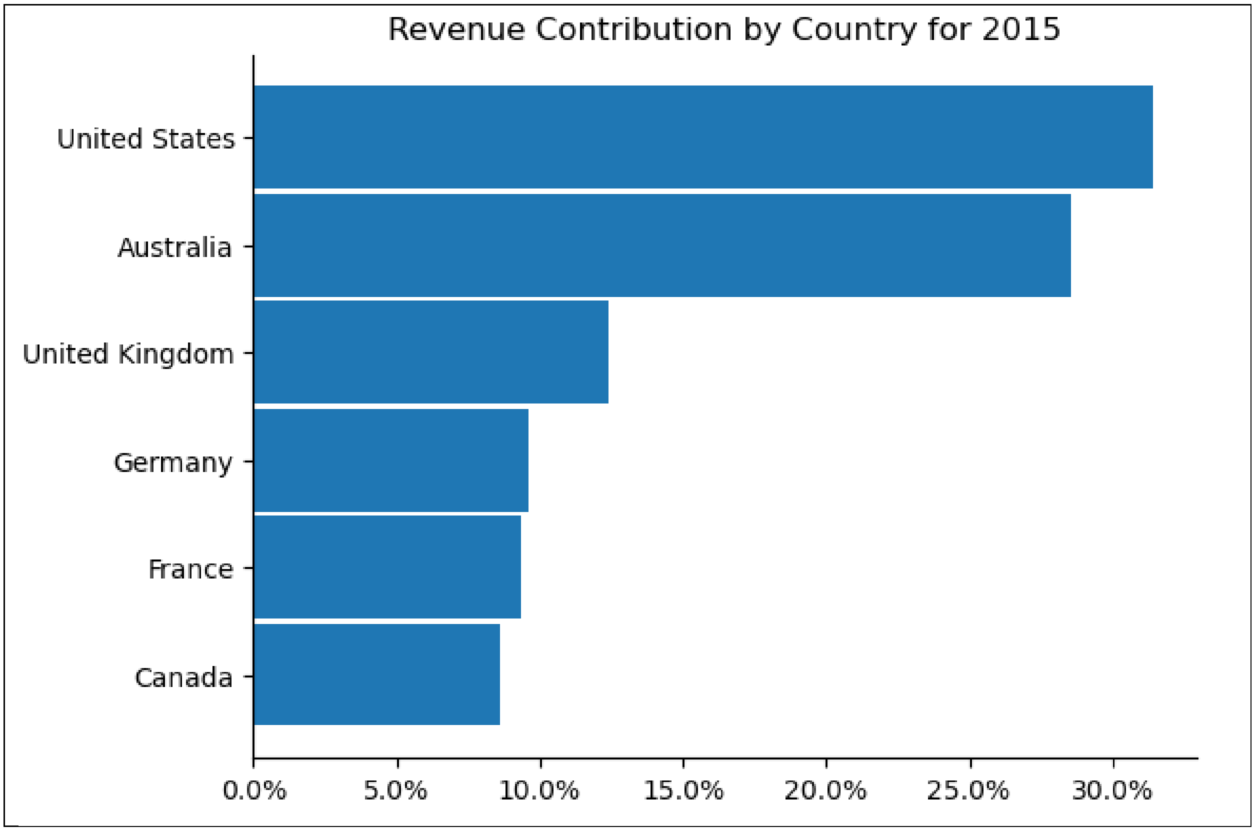
Looking at all the data available, let's plot the customer age and the number of orders received. We can quickly see that, male or female, customers aged 30-40 have placed the most orders. Not surprisingly, the number of bicycle product orders diminishes as customers get older.



Questions to explore further:

- What products are people buying in the various age categories?
- Which age categories are the most, and least, profitable?

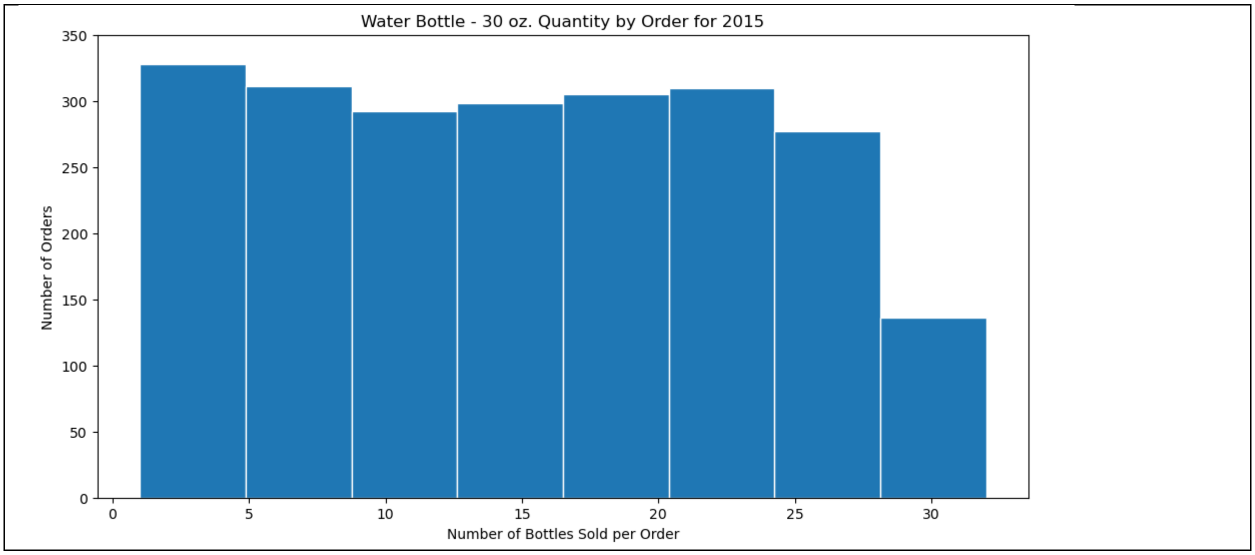
How does each country’s revenue contribute to overall revenue in 2015?



Questions to explore further:

- We would most certainly like to know more about what products are selling in each country.

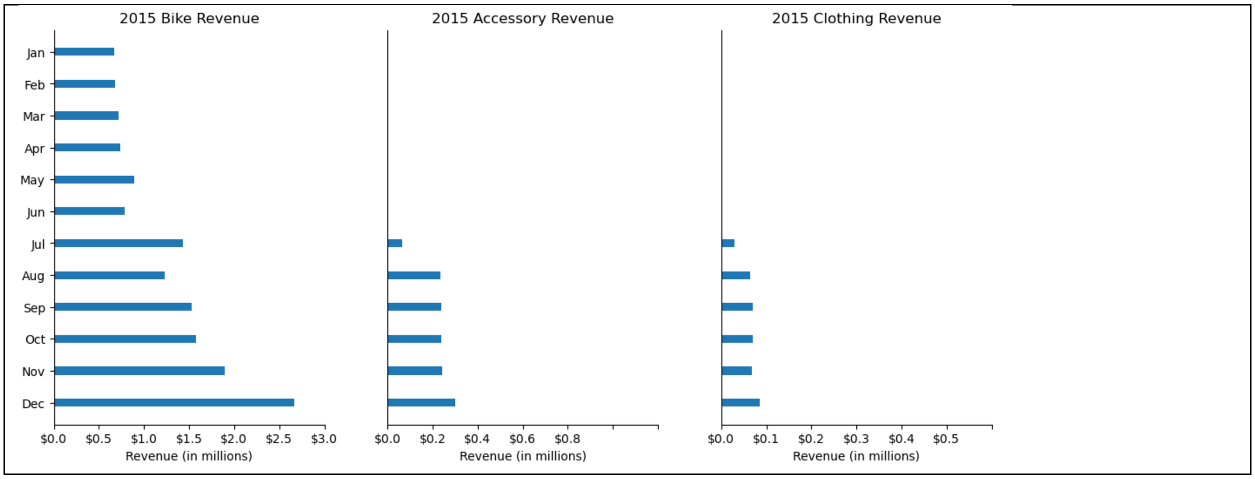
In 2015, nearly 2300 orders were placed for 30oz Water Bottles, and generally a few were ordered at a time. Most orders were for less than 5 bottles, however other order amounts were common too.



Questions to explore further:

- Are there volume discounts for water bottles?
- Do people regularly loose water bottles, and need to replace them?

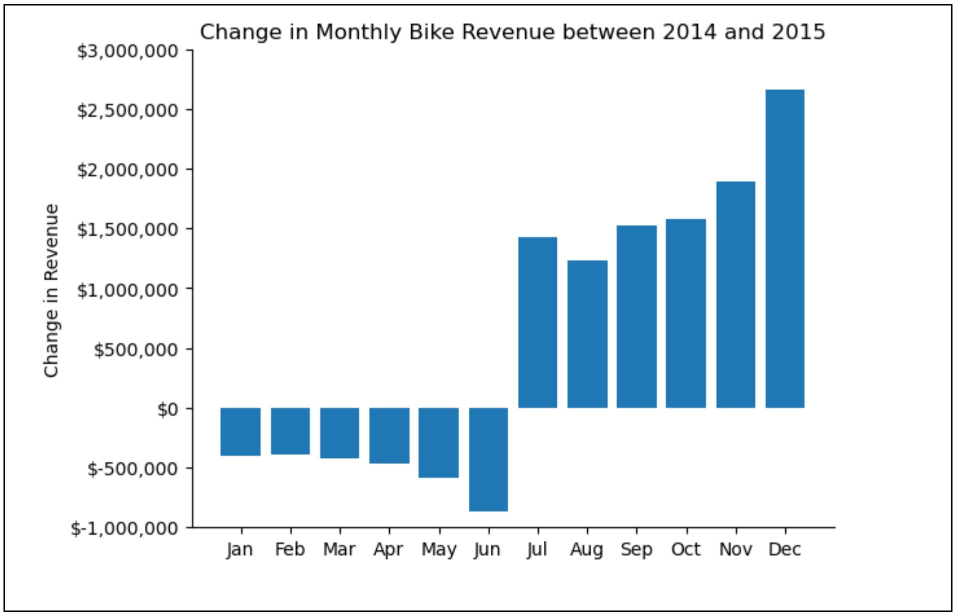
I looked further at 2015 revenue data to see if there might be any correlation between Product Categories. For example, you might expect accessory and clothing sales to increase as more bikes are sold. Since people tend to bicycle more in the warmer months, you might expect sales to increase at those times.



Looking at data, raises several questions:

- Why were there no sales of accessories or clothing during the first six months of the year? Are we missing data?
- The increase in sales during the fall and early winter months is unexpected too. We would need to know more about the business to answer.

Let's compare the change in monthly bike revenues between 2014 and 2015.

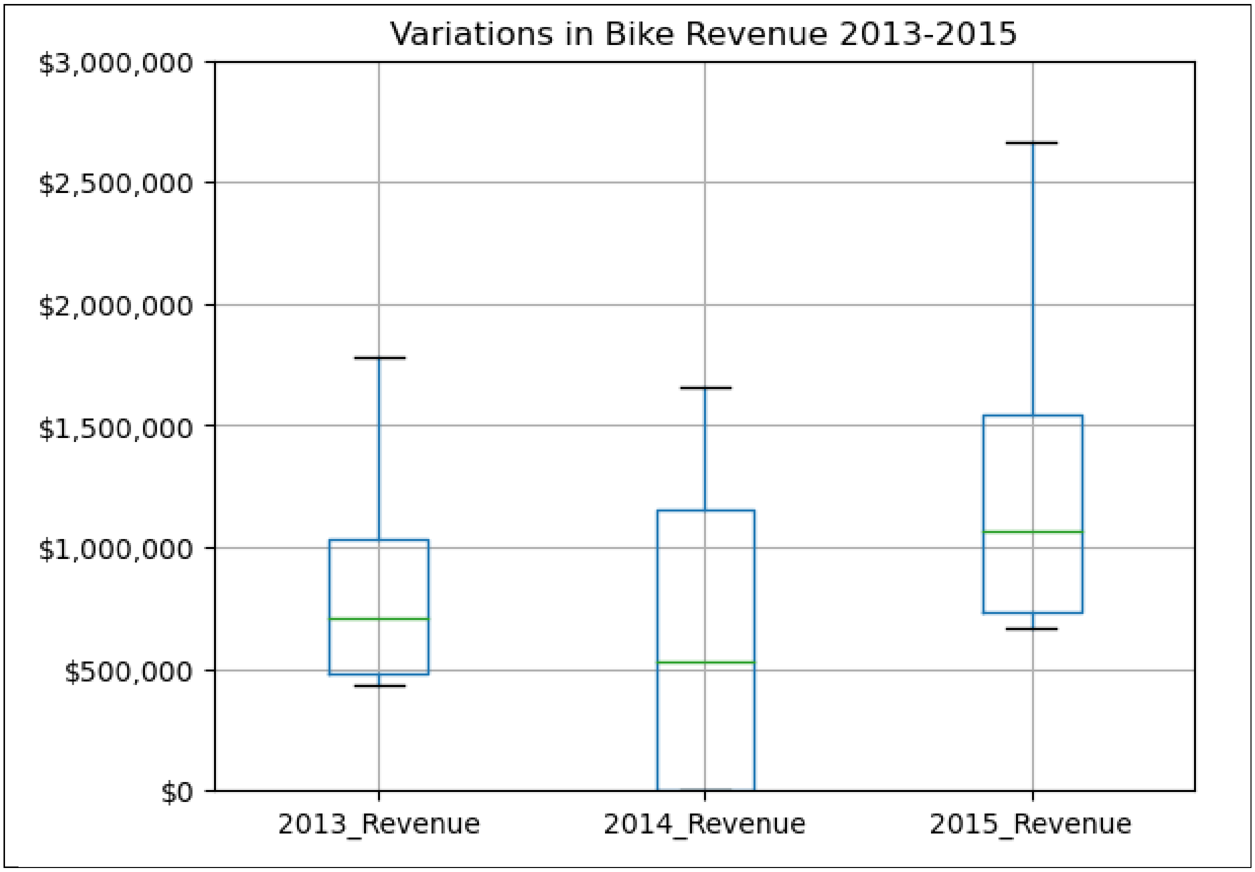


Bike revenue does not look good for the first six months of year, with each month in 2015 down from 2014. The second half though looks great. However, let's look at the data.

	Bike Revenue		
Month	2014	2015	Change
Jan	$1,068,475	$666,938	($401,537)
Feb	$1,065,852	$679,246	($386,606)
Mar	$1,135,880	$714,008	($421,872)
Apr	$1,202,496	$734,608	($467,888)
May	$1,482,909	$895,043	($587,866)
Jun	$1,655,631	$786,480	($869,151)
Jul	$0	$1,427,219	$1,427,219
Aug	$0	$1,232,390	$1,232,390
Sep	$0	$1,527,727	$1,527,727
Oct	$0	$1,579,537	$1,579,537
Nov	$0	$1,890,854	$1,890,854
Dec	$0	$2,665,033	$2,665,033
Totals	$7,611,243	$14,799,083	$7,187,840

We're surely missing bike revenue data for 2014 and will need to investigate this.

And lastly, let's compare Bike Revenues for 2013-2015.



Again, we can see the effects of missing data from 2014. The overall revenue trend is rising with more monthly variation.

	Bike Revenue		
Month	2013	2014	2015
Jan	$435,063	$1,068,475	$666,938
Feb	$485,405	$1,065,852	$679,246
Mar	$466,940	$1,135,880	$714,008
Apr	$476,905	$1,202,496	$734,608
May	$589,689	$1,482,909	$895,043
Jun	$534,099	$1,655,631	$786,480
Jul	$924,632	$0	$1,427,219
Aug	$827,458	$0	$1,232,390
Sep	$1,021,403	$0	$1,527,727
Oct	$1,061,864	$0	$1,579,537
Nov	$1,257,785	$0	$1,890,854
Dec	$1,777,544	$0	$2,665,033
Totals	$9,858,787	$7,611,243	$14,799,083