Bike Sales in Excel- Data Cleaning and Data Visualization

Snippet of the bike sales data:

ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Dista	Region	Age	Purchased Bike
12496	M	F	$40,000.00	1	Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	42	No
24107	M	M	$30,000.00	3	Partial College	Clerical	Yes	1	0-1 Miles	Europe	43	No
14177	M	M	$80,000.00	5	Partial College	Professional	No	2	2-5 Miles	Europe	60	No
24381	S	M	$70,000.00	0	Bachelors	Professional	Yes	1	5-10 Miles	Pacific	41	Yes
25597	S	M	$30,000.00	0	Bachelors	Clerical	No	0	0-1 Miles	Europe	36	Yes
13507	M	F	$10,000.00	2	Partial College	Manual	Yes	0	1-2 Miles	Europe	50	No

Here is what the data looks like after data cleaning:

ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	Age Range	Purchased Bike
12496	Married	Female	$40,000	1	Bachelors	Skilled Manual	Yes	0	0-1 miles	Europe	42	Middle Age	No
24107	Married	Male	$30,000	3	Partial College	Clerical	Yes	1	0-1 miles	Europe	43	Middle Age	No
14177	Married	Male	$80,000	5	Partial College	Professional	No	2	2-5 miles	Europe	60	Old	No
24381	Single	Male	$70,000	0	Bachelors	Professional	Yes	1	5-10 miles	Pacific	41	Middle Age	Yes
25597	Single	Male	$30,000	0	Bachelors	Clerical	No	0	0-1 miles	Europe	36	Middle Age	Yes
13507	Married	Female	$10,000	2	Partial College	Manual	Yes	0	1-2 miles	Europe	50	Middle Age	No

Steps to clean the dataset and start drawing our analysis:

1. **Checking for duplicates:** The dataset that we worked on had 26 items that were duplicates. Duplicate items in a dataset tend to throw off the analysis, and therefore getting rid of it is the first priority.
2. **Changing the column values to avoid confusion:** From the above image, what we did first is we cleaned up a bunch of columns like 'Marital Status' , 'Gender', Income, and Age. For age, we added another column (Age Range) to distinguish the age and break that using nested ifs. The new column has values- "Adolescent", "Middle Age", and "Old". We also replaced the marital status column values with 'Married' and 'Single'. We then added the Gender column values as 'Female' and 'Male'. This avoids confusion between the "M's" for the two columns.
3. **Getting rid of extra decimals for the Income column:** since in the dataset, the income column is of the datatype currency. We keep the datatype as currency and got rid of the extra 2 digit decimal value to make our data look more clean.

Building out dashboards:

To build out dashboards, we made use of **Pivot tables.**

1. **Average Income per Purchase:** we analyzed the average income amongst the two genders and drew the output as to which gender has the most income and purchased the bike.
2. **Purchased Bike per Age:** This line chart shows the age range among the customers who purchased the bike.
3. **Purchased Bike per Car Owned:** The bar chart shows the number of cars owned per customer and their need to purchase a bike or not.
4. **Distance covered after Purchasing the bike:** We saw a downward trend where if the distance to commute is more than 5 miles, the number of purchase is lower. 200 of the customers although did buy a bike to maybe run arround some errand or just do some cardio
5. **Occupational Customers buying a Bike:** From the data given we see, Professionals usually opted on purchasing the bike. And if you are someone like me, I would love to take my bike to work in the summers.