Excel Project

Introduction

This project is the steps I took to clean and analyze data from a bike company. The attached Excel spreadsheet contains sheets for the raw data, the working data I used to clean, the pivot tables and charts I created, and the final dashboard. Below, I have my process broken into steps within 4 larger categories – Cleaning, Analyzing, Building the Dashboard, and finally, the insights. This is meant as a total look into my process for using Excel to complete a data analysis project from start to finish. It can be followed step by step with the attached data, or you can skip down to the insights to see some interesting stats I uncovered in this dataset.

1. Take a look at data.

* Bike sales with demographic information about the bike buyers/non-buyers.
* Unique id number, gender, income, education status, kids, cars, etc.
* At this point, I selected all data and added filters to each column header. Checked numeric columns to ensure no 0s or blanks where they don’t make sense.

Cleaning

1. Remove duplicates.

* There were 24 duplicates which Excel removed. Now 1001 unique values remain

1. Change the columns that have initials rather than words to make data more readable and usable.

* Under Martial Status column, M and S changed to Married and Single. Use Control + H, or Find and Replace feature, to change those.
* Same as above for Gender column, with M being Male and F being Female.

1. Income Column – upon inspection, there are no decimal values, so I changed the data type to currency and went out 0 decimal places to make data look cleaner/simpler.

* If I were going to be doing in depth calculations with this data, I would reformat to Number data type to prevent errors.

1. Children, Education, Occupation, Homeowner and Cars columns all look good when examining on filter.

* No spelling errors, no blanks or inappropriate values within each column. Used the filter to examine.

1. Commute Distance column is a range within each cell, measured in miles. This looks fine for now, but if I come across an issue when create pivot tables or visualizing, I can always update.
2. Region Column looks good.
3. Age column

* There are so many ages, and when I visualize it, it might look muddled. The range goes from 25-89, so I’m going to create mini-ranges or brackets of ages to group individual ages in.
* Create new column, Age Bracket. Use IF function to categorize the ages.
* =IF([cell reference] < 31, “Adolescent”, “Invalid”)
* This works, so now I’m going to build on that and make a nested IF statement.
* =IF([cell reference] >= 31, “Middle Age”, IF([cell reference] < 31, “Adolescent”, “Invalid”))
* Now I’m going to break it even further and make above the value IF FALSE.
* IF([cell reference] > 54, “Old”, IF([cell reference] >= 31, “Middle Age”, IF([cell reference] < 31, “Adolescent”, “Invalid”)))
* Now I have 3 distinct age brackets, 30 & under, 31- 54, and 55+, as Adolescent, Middle Age and Old respectively which is much more usable than just the age. I chose these names for each bracket so that when I visualize the data, the brackets will appear in ascending order of age based on the ascending alphabetical order.

1. Purchased Bike column looks great.
2. That concludes cleaning. There wasn’t a tremendous amount to clean.

Analyzing

1. Create some pivot tables, from which I’ll build a dashboard.
2. First Pivot Table: Average Income of customers, with gender and bike purchase included.

* I did reformat the income here to show only 2 decimal places, rather than the 4 it was auto-populating.
* From this, I can see that the average income of males is higher than females, and the average income of bike purchasers is higher than non-purchasers.
* From this first pivot table, I created a Pivot Chart. Double bar chart, with gender on the x-axis, colored bars representing bike purchase, and average income on the y-axis. I reformatted the income data type to look neater on the chart.
* Housekeeping on the chart like giving each axis a title and naming the overall chart was done, and the chart is ready to go.

1. Second Pivot Table: Looking at commuting distance to work and see if that impacts if the customer purchased a bike or not.

* I created a Pivot Table with commute distance, a count of bike purchases, and the bike purchase data (yes or no).
* Upon reviewing this Pivot Table, I can see that the 10+ miles bracket is out of order, because of how Excel is reading the number -- the plus sign is within the string of the number 10. This skews the table, and it will skew the visualization because that particular bracket is the largest commute and should therefore be after the 5-10 mile bracket. I went back into the sheet with my working data and used Find and Replace to replace “10+ miles” with “10 + miles” instead, so that Excel can read the number 10 correctly and order it. This alleviated that problem.
* I can see that the highest bike purchase grouping is also the shortest commute distance, which is expected. The lowest bike purchase grouping is also the highest commute distance. The only unexpected finding is that the 3 middle commute distance groups did not quite decrease in bike purchases as commute distance increased; for example, the second highest number of bike purchases belonged to the middle category of commute distance, 2-5 miles.
* From the Pivot Table I created a Pivot Chart using a line chart. This chart displays the above findings.
* I cleaned up my chart by adding axis labels, as well as a chart title.

1. Last Pivot Table: Customer Age Brackets

* I created a Pivot Table with the age brackets as columns, the count of bikes purchased as a value, and bike purchase information as rows. This shows that the age bracket looking at bikes the most by a long shot is the middle age bracket, or ages 31-54. Those under 31 and over 54 are not really even looking at bikes, they are making up the smallest portions of our customer base. Also, while there weren’t many more middle age customers buying bikes versus not buying bikes, they were the only age bracket of the 3 to have a larger portion buying bikes than not. Both adolescent and old ages were not buying bikes more than they were buying bikes.
* I then created a line chart with this data, with the bike purchase information as the legend, the count of bikes purchased as the y-axis and the age brackets as the x-axis. This type of chart best illustrates the difference in bike customers by age bracket.
* I cleaned up my chart by adding a chart title, as well as axis titles.

Building the Dashboard

1. Now, I can copy and paste my Pivot Charts into a new sheet to build this dashboard.
2. I’m going to remove the gridlines from the actual spreadsheet itself to make this look cleaner.
3. I added a title/header with color to make it look nice and clean.
4. I situated the charts how I wanted them to look in terms of placement. I placed commute distance chart at the bottom of the other 2 charts because it’s a longer chart, and I formatted it so that it spanned the length of the top 2 charts together.
5. Now, in order to add interesting filters to the dashboard, I’m going to insert slicers which will be filter menus for the dashboard.

* I added slicers for Marital Status, Education and Region. I also reformatted the size of each to ensure they aren’t wasting space and look nice. To apply these filters to all 3 charts, I edited the report connections in the slicer menu to include all 3 pivot tables I created.
* With these slicers, I can filter the data and pull interesting information to gain insights for the bike company, such as average income in North America for those who have Partial College as their education level, or how many middle age people in Europe who commute 2-5 miles to work are purchasing bikes.

1. This concludes the dashboard building portion of this project.

Insights

1. Men are on average making between $3000-$4000 more than women in income annually.

* Marriage only increases these numbers. Both married men and women make more income annually.
* The average income of men who purchase a bike is around $60,000. The average income of men who don’t purchase a bike is around $56,000. This is a fairly slim margin, however if the bike company wants to increase bike sales, they may consider lowering the cost of the bikes, or running a promotion marketed to this particular income segment to gain more purchases. Without the bike price data, it’s unclear if this would negatively impact the bottom line, by decreasing bike prices, but it’s an option.
* Targeted ads to women might increase the bike purchases for this demographic. Also consider product designs to possibly move the needle on female bike purchases.
* Since the income is higher for both married men and women, especially those who purchase bikes, an option moving forward could be to produce a tandem bike option for couples at a higher price point than regular bikes. Because married men and women have a higher average annual income, especially of those who purchase bikes, they might be interested in purchasing a bike they can ride together, and they can afford a higher price point. I would recommend sending out customer surveys, including questions to gauge interest in tandem bikes for married couples and also questions regarding income. This will help determine if a tandem bike would be a worthwhile product to produce. If not, another option might be some sort of “buy one get one” at a lesser rate for couples, in the style of “his and hers bikes.”
* Middle aged married people purchased 180 bikes, whereas single middle aged people purchased 200 bikes. The above suggestions might increase sales for married middle aged people. This is a key demographic, as married men and women have a higher income, and middle age represents the highest segment of customers. If the bike company could increase bikes purchased by married middle aged people, these would likely be better customers who have more disposable income to purchase higher ticket bikes, more bike accessories, etc.

1. The under 31 age bracket has the lowest representation of customers. This age bracket has purchased less than 50 bikes, and they more often do not buy at all.

* Roughly half of the bikes purchased by this age bracket were from those in the Pacific region. This was the only region to have more under 31s buying bikes than not. The Pacific region does make up the smallest segment of the customer base. An option might be to expand in this area, to increase their representation of the base, as both adolescent and middle age customers are more likely to buy bikes than not in this region.

1. Those customers with Bachelors and Graduate degrees are purchasing more bikes than those with other education statuses.

* These groups also have a higher annual income than other groups because of their higher education level and make up more of the middle aged bracket. This is a group the bike company should consider as prized customers and cater marketing and advertising to.
* An option to increase visibility with this group would be to increase advertising on college campuses. The bike company could even set up booths and/or pass out fliers or coupons on campus.

1. The commute distance from work data has some puzzling information.

* While it’s not surprising that those who work closest to home, 0-1 miles, buy more bikes, the next highest category is those who work 2-5 miles from work. Those who work 1-2 miles from home purchase less bikes than those who work 2-5 miles away. It isn’t surprising that the further 2 distance brackets purchase less bikes.
* It might be worthwhile to conduct customer surveys for those who work 0-5 miles from home, asking questions about modes of transportation, work times and non-work activities/hobbies. This sort of survey will help the bike company determine if there are other factors at such as odd commute times for bike-riding, other preferred transportation like walking or driving, or other sources of exercise such as gym memberships at nearby/office gyms. With the insights gained from this survey, the bike company can make decisions about marketing or advertising campaigns directed at low-distance commuters.

1. I added in a slicer for Occupation data after looking at commute distance data above, because I wanted to see which specific occupations purchased more bikes, made a higher income, and where they fell in terms of commute distance.

* Overall, those who work in management and professional fields make the most income out of the 5 occupation categories. Those who work in management average between about $85,000 and $88,000, and those who work in professional fields average between about $74,000 and $75,000. In both men and women, professional and management occupations, the average income of bike purchasers is higher than non-bike purchasers.
* There is a discrepancy, however, in the count of bikes purchased by occupation. Those in management are purchasing significantly less bikes, regardless of commute distance. In fact, the only commute distance which saw a higher number of bikes sold to management workers was 0-1 miles, and there were less than 5 purchases separating those who did and didn’t buy bikes. They also purchase less bikes as they age in the management field. This particular demographic is worth digging into, because these customers have a higher average income by about $10,000, and would potentially be good customers for the bike company. I recommend the bike company conduct a customer survey of those in management to determine their length of workday, work-life balance, activity levels and desire for exercise; these types of questions will lead to data that the bike company can use to target this high-income group and make a larger impact on their bottom line.