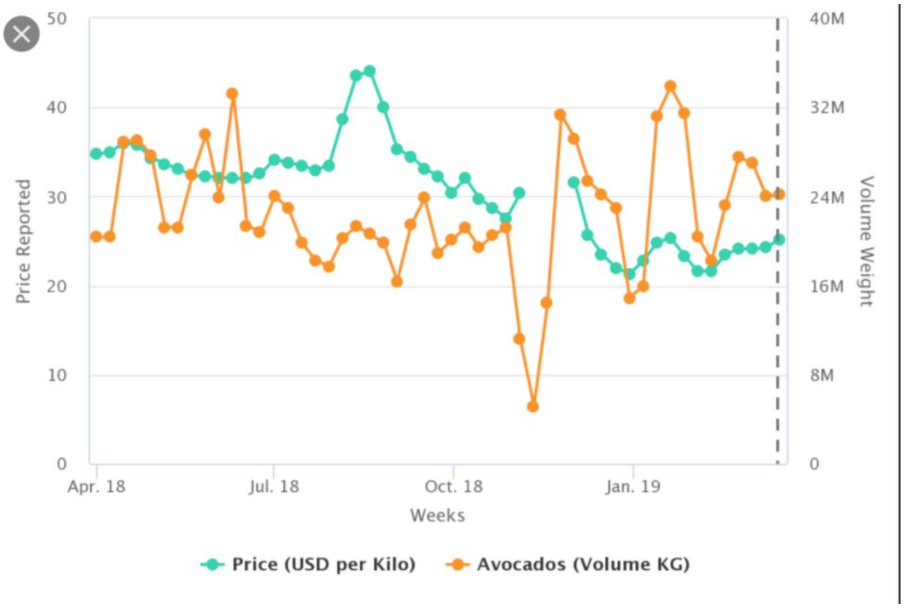
EDA and Visualization: Final Project

Sophie Wang and Christabelle Pabalan

Visualization One:

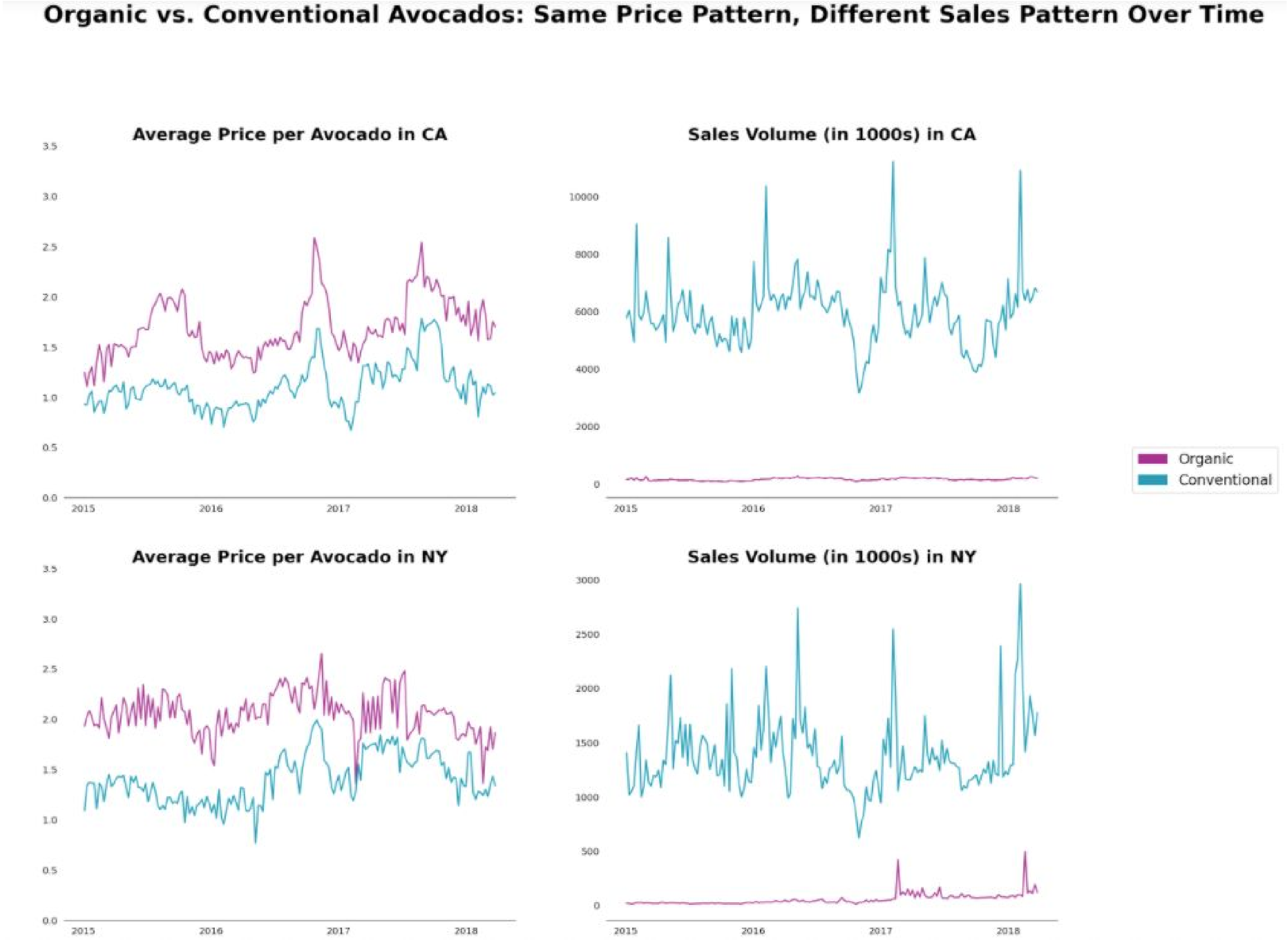
Original Plot:



Identify the Story:

The original plot is showing the trend of avocado price and sale volume in the US from April 2018 and January 2019. The blue line represents the price changes and the orange line represents the sales volume changes. We can see that over this one-year period, avocado price decreases while sales volume remains at a similar level with some fluctuations over the period.

Improved Visualization:



Our Changes:

1. **Removed the double axes and muted the grid line.** The original plot has a double axis to show price and sales volume of avocado. Double axes make it more difficult to interpret and understand the trend line. In our plot, we leverage subplots to separate the visualizations showing price trend and sales volume trend. We also muted the grid line who showed up in the original plot. The grid line does not add value, instead it makes the plot look cluttered.

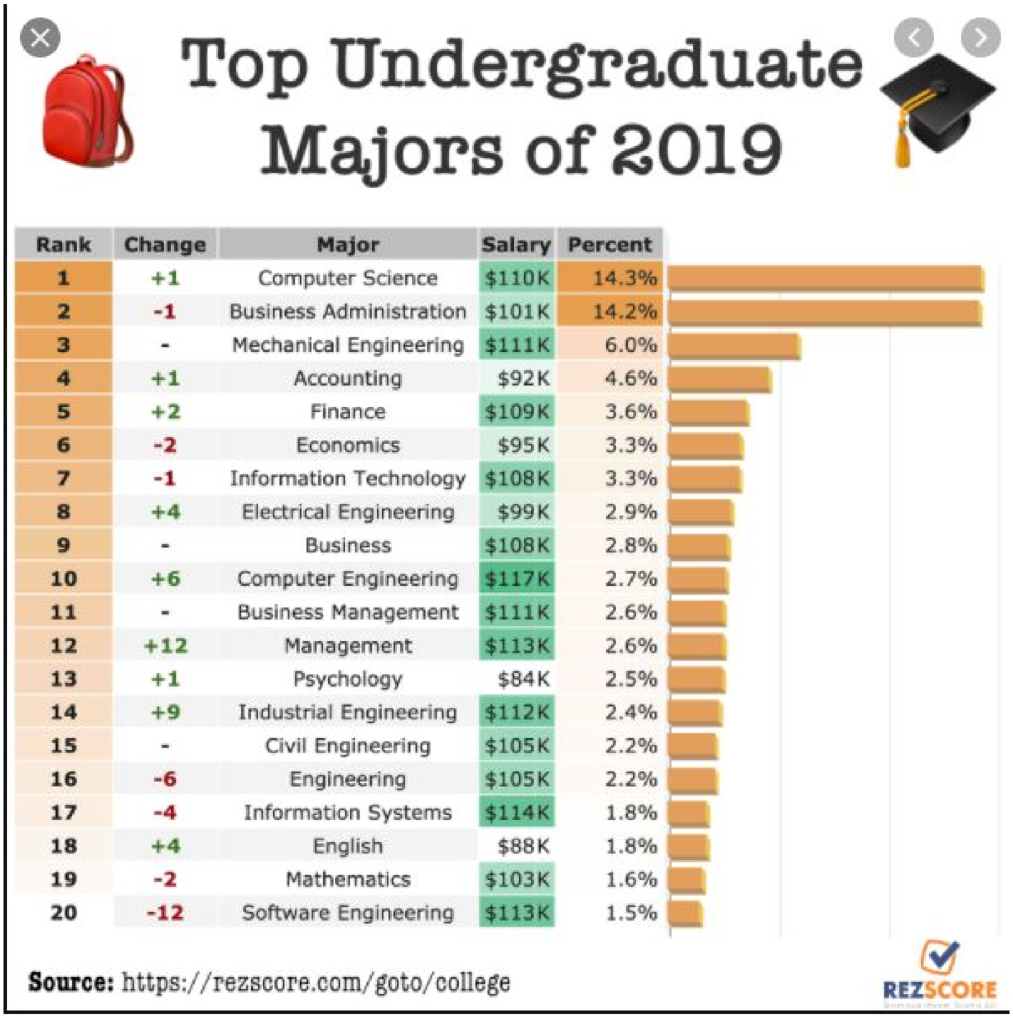
2. **Removed the markers on the trend line.** The markers on the original plot are big and distracting and did not add any values. We did not add markers on the trend line.
3. **Made all the text horizontal.** In the original plot, the y labels are vertical, which is harder to read. In our plot, we made sure no horizontal text appears.
4. **Used subplots to show more stories**. We found a dataset which shows the price per avocado and total sale volume for different states in the US from 2015 and 2018. We chose to narrow down the scope to California and New York only. The dataset also includes the information for organic versus conventional avocados. So we can show the trend differences for organic versus conventional avocados too.

Story from the Improved Visualization:

1. In both CA and NY, the price per organic avocado is higher than conventional avocado (as expected). The prices are relatively stable over the period from 2015 and 2018 in two states. However in CA, there are two noticeable peaks for both organic and conventional around the end of 2016 and 2017, but after the peak the price for each avocado falls back to the previous level again.
2. By looking at the two plots “Average Price per Avocado in CA” and “Average Price per Avocado in NY” , we can see that even though the price trends are similar in CA and NY, unit price for avocado in NY is slightly higher than in CA. Maybe due to more agricultural areas in California Central Valley. So the supply of avocado in California is higher and the shipment cost is lower.
3. For both CA and NY, even though the price trends of organic and conventional avocados are similar, the sales volume trends between organic and conventional avocados are different. Over the period of 2015-2018, the sales volume for conventional avocado fluctuates a lot and much higher than that for organic avocado, whereas the sales volume for organic avocado remains low and smooth.

Visualization Two:

Original Plot:



Top Undergraduate Majors of 2019

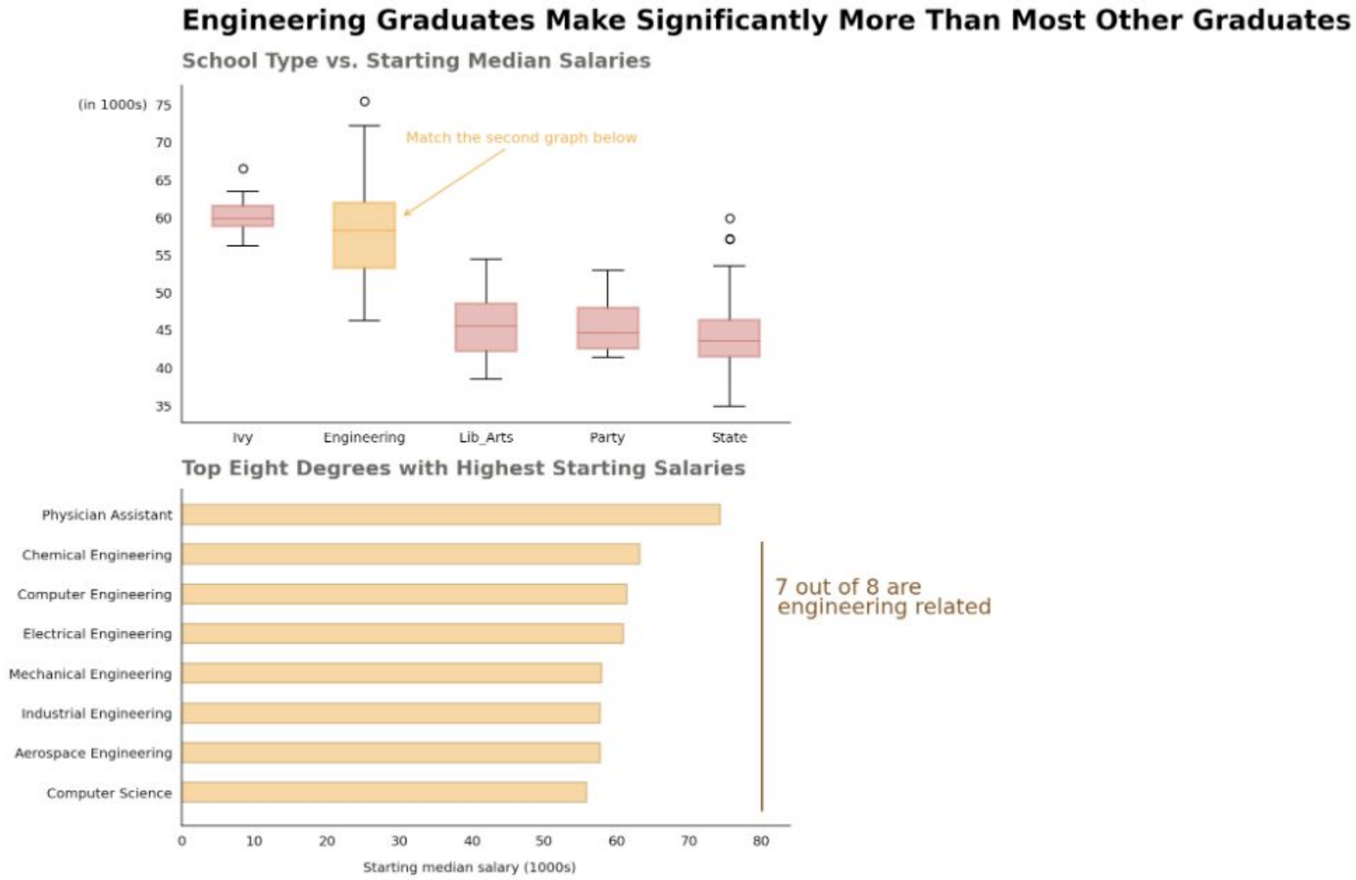
Rank	Change	Major	Salary	Percent
1	+1	Computer Science	$110K	14.3%
2	-1	Business Administration	$101K	14.2%
3	-	Mechanical Engineering	$111K	6.0%
4	+1	Accounting	$92K	4.6%
5	+2	Finance	$109K	3.6%
6	-2	Economics	$95K	3.3%
7	-1	Information Technology	$108K	3.3%
8	+4	Electrical Engineering	$99K	2.9%
9	-	Business	$108K	2.8%
10	+6	Computer Engineering	$117K	2.7%
11	-	Business Management	$111K	2.6%
12	+12	Management	$113K	2.6%
13	+1	Psychology	$84K	2.5%
14	+9	Industrial Engineering	$112K	2.4%
15	-	Civil Engineering	$105K	2.2%
16	-6	Engineering	$105K	2.2%
17	-4	Information Systems	$114K	1.8%
18	+4	English	$88K	1.8%
19	-2	Mathematics	$103K	1.6%
20	-12	Software Engineering	$113K	1.5%

Source: https://rezscore.com/goto/college

Identify the Story:

The original plot is using a horizontal bar chart to show the ranking of salaries by undergraduate majors.

Improved Visualization:



Our Changes:

1. **Removed all the decorations**. In the original plot, there are multiple stickers (backpack and graduation hat) which is very distracting and does not add any values. We remove those unnecessary decorations.
2. **Removed the unnecessary colormap.** In the original plot, it uses different shades of grey on the ranking numbers, which is absolutely meaningless. In the improved plot, we did not add any unnecessary colormap.
3. **Restructure the visualization**. We understand that the purpose of the original plot is to show the major and corresponding starting salaries with descending order. But the original plot does not look professional with too much clutter so we decided to redesign it with different plots. We found two relevant datasets which include the school types vs. starting median salaries and undergraduate majors vs. starting median salaries. So we decided to use a boxplot to show the distribution of starting salaries for each school type (Ivy League, Engineering, Liberal Arts, Party, State) and used a horizontal bar chart

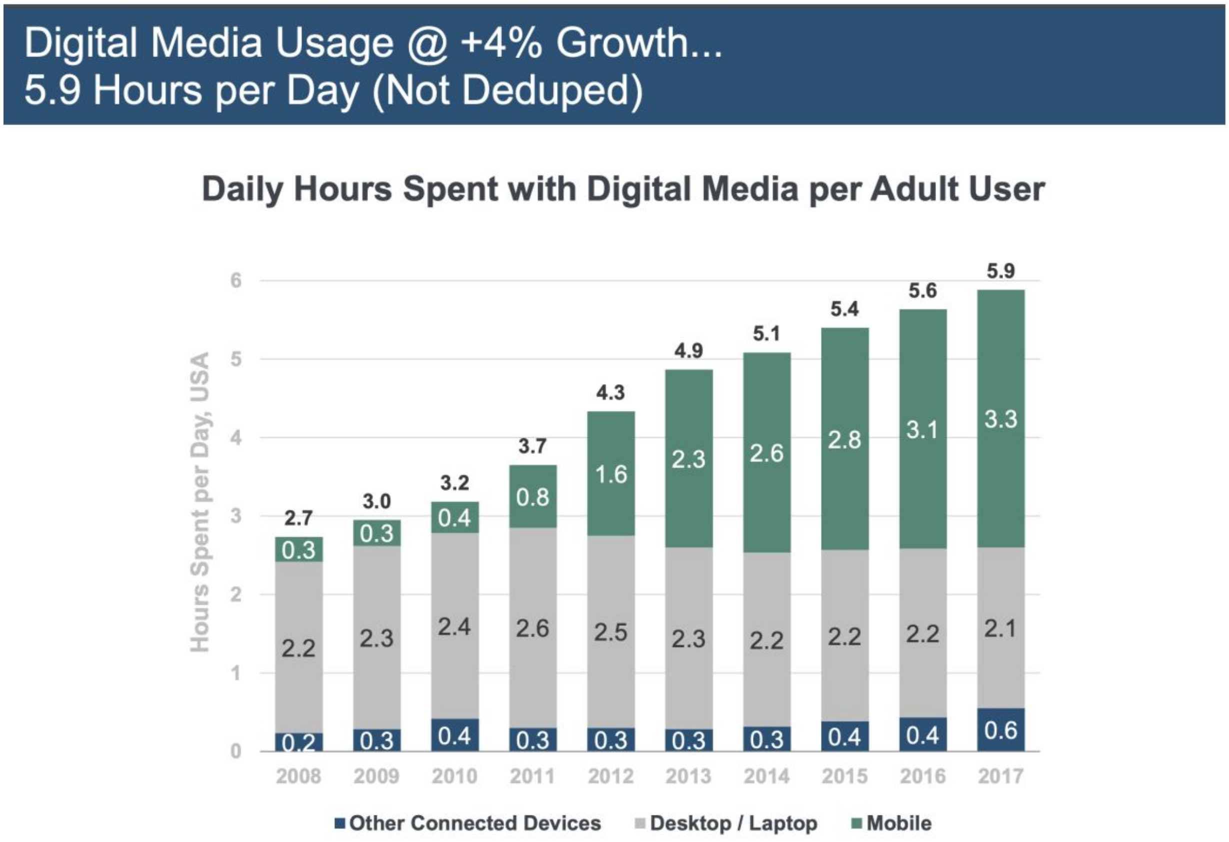
to the top eight lucrative undergraduate majors and their starting median salaries. In the box plot, we found out that even though Ivy League graduates have the highest median starting salaries, engineering school graduates have much higher upper limit starting median salaries. It means, people with the highest starting salaries are most likely coming from engineering schools like MIT/Caltech, instead of Ivy Leagues. It totally matches the fact that seven out of eight highest starting median salaries are from engineering related majors, as shown in the horizontal bar chart. So we **used the similarity principle** and marked the engineering school boxplot with the same color orange with the top eight highest paying majors bar chart's color.

Story from the Improved Visualization:

1. In the box plot, we found out that even though Ivy League graduates have the highest median starting salaries, engineering school graduates have a much higher upper limit of starting median salaries.
2. This means, people with the highest starting salaries are most likely coming from engineering schools like MIT/Caltech, instead of from Ivy Leagues. It totally matches the fact that seven out of eight highest starting median salaries are from engineering related majors, as shown in the horizontal bar chart.
3. By school type, Liberal Arts, Party and States schools have similar starting median salaries, about $15,000 lower than Ivy League and Engineering schools. While Ivy League and Engineering schools focus on engineering and STEM majors, Liberal Arts, Party and States usually have many graduates who study liberal arts majors which tend to generate less salaries.
4. Although it's totally a personal preference of choosing majors, from purely salaries point of view, the best option is to go to either Ivy League and engineering school and study one of the engineering majors based on the visualization shown.

Visualization Three:

Original Plot:



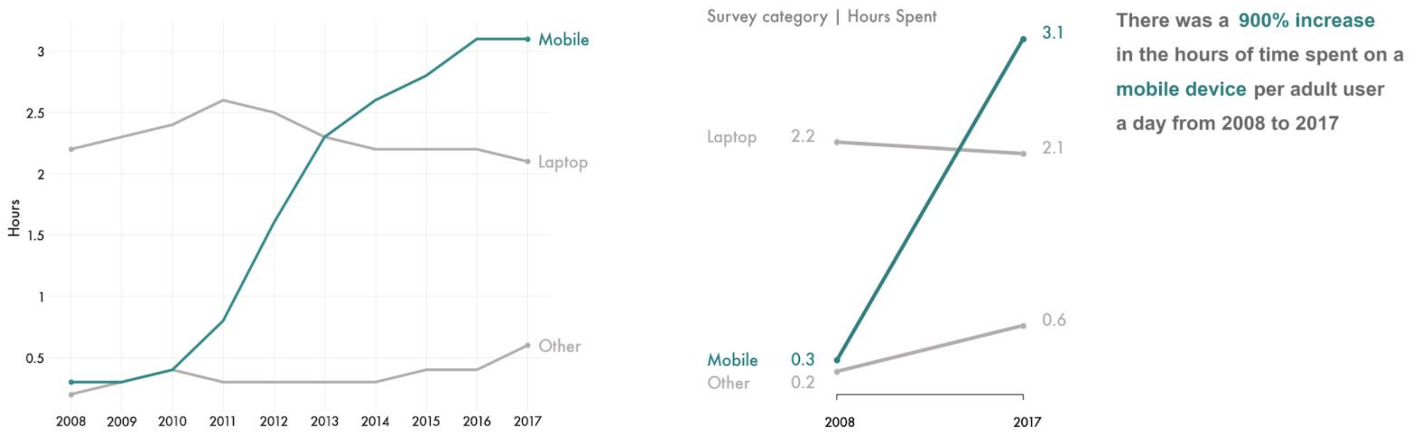
Digital Media Usage @ +4% Growth...
5.9 Hours per Day (Not Deduped)

Daily Hours Spent with Digital Media per Adult User

Identify the Story:

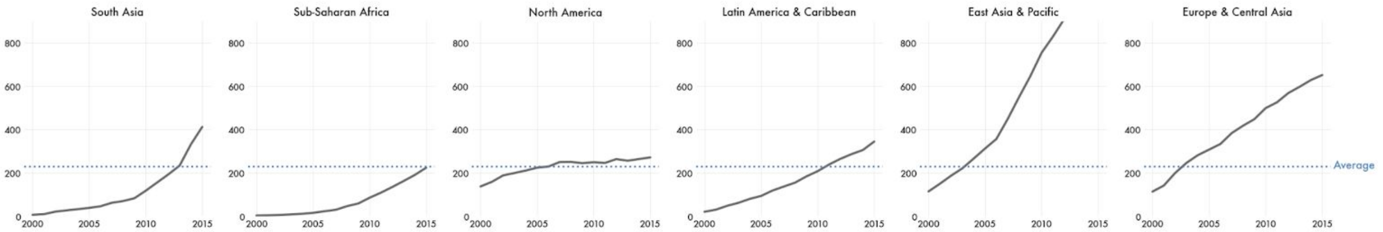
The original plot reveals the increase in the overall daily hours spent on digital media per adult user throughout the years. We can see that our overall time spent on digital media increases with time. Furthermore, we can also see that the hours spent on mobile devices appear to be the only category that is consistently increasing with time as seen in the expansion of the green bars.

Improved Visualization:


Time Spent per Adult User a Day with Digital Media in the US, 2008 to 2017

There was a dramatic increase in the hours of time spent on a mobile device compared to the laptop or other categories which remained relatively the same

There was a **900% increase** in the hours of time spent on a **mobile device** per adult user a day from 2008 to 2017


Number of Internet Users by World Region (in millions), 2000 to 2015

Our Changes:

1. We synthesized data from our world in data and kleinerperkins in order to show both the progression of time spent on each type of digital device and the amount of internet users throughout the years.
2. The previous graph had competing colors and, as a result, the mobile category didn't stand out immediately. We decided on decreasing the cognitive load of the audience by **making use of contrast.** Everything else in the graph is greyed out except for the teal highlighted objects which allows these objects to pop out at the audience.
3. Additionally, we used the **similarity principle** by making the objects referring to the "mobile category" colored teal. This is to immediately indicate that these objects belong to the same group across the entire graph.
4. We also changed the stacked bar chart into multiple line graphs in order to display how each category has changed throughout the years independently. This is also intended to

further highlight the substantial increase in the time spent on a mobile device **by contrasting against the other lines that have remained relatively consistent.**

5. We decided on including a slopegraph which includes the end point years in order to display the extent of the increase from 2008 to 2017 more clearly.
6. Utilized the **proximity principle by directly labeling our graphs.** This helps our audience group together the line and the category immediately.
7. Included a plot on the number of internet users by world region across time from 2000 to 2015
8. Omitted the distracting border and decreased overall clutter

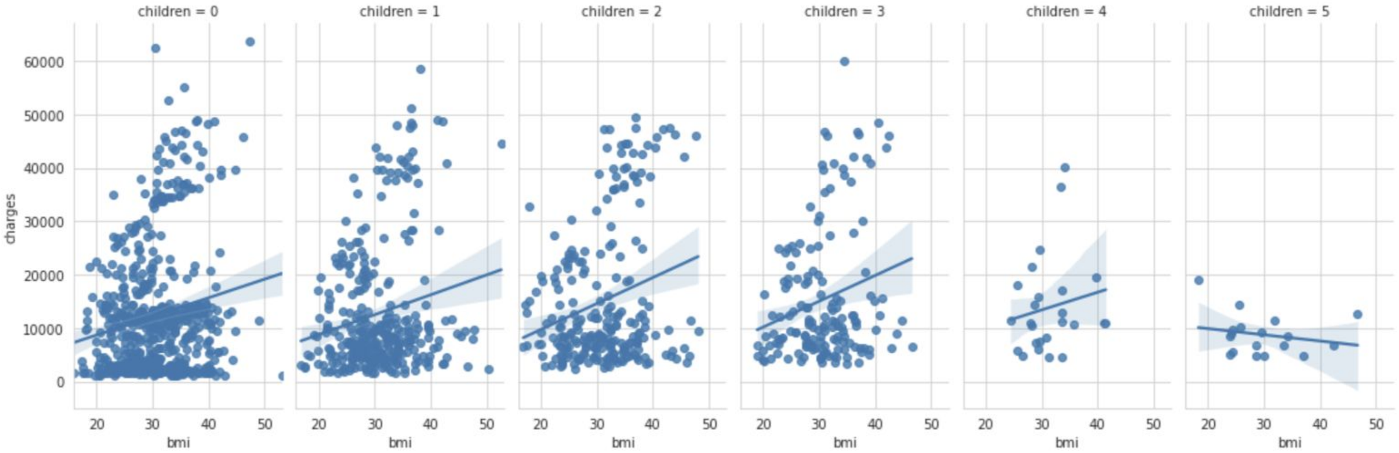
Story from the Improved Visualization:

1. The story from the previous visualization is more centered around the collective increase in hours spent on digital media.
2. Our revamped visualization is heavily **highlighting the increase of time spent on our mobile devices** and comparing this to the baseline of the increase of our time spent on a laptop or other devices.
3. Our time spent on our mobile devices has increased from 0.3 to 3.1 hours on average from 2008 to 2017.
4. On the slopegraph, we can see **a sharp increase in mobile devices and a slight decrease in our time spent on our laptops.** This decrease can be attributed to the distributed time now spent on our mobile devices. However, this very slight slope down in the slopegraph could be due to variance since the time spent on our laptops have been relatively stable.
5. This hike in mobile device usage is a reflection on the extent in which the evolution of handheld devices grew in sophistication, in adaptation and in popularity.

6. We also added an extra component that highlights the number of internet users in each world region in the millions. Even at that large of a scale, we could see that as a function of time, **the number of internet users across all world regions have increased.**

Visualization Four:

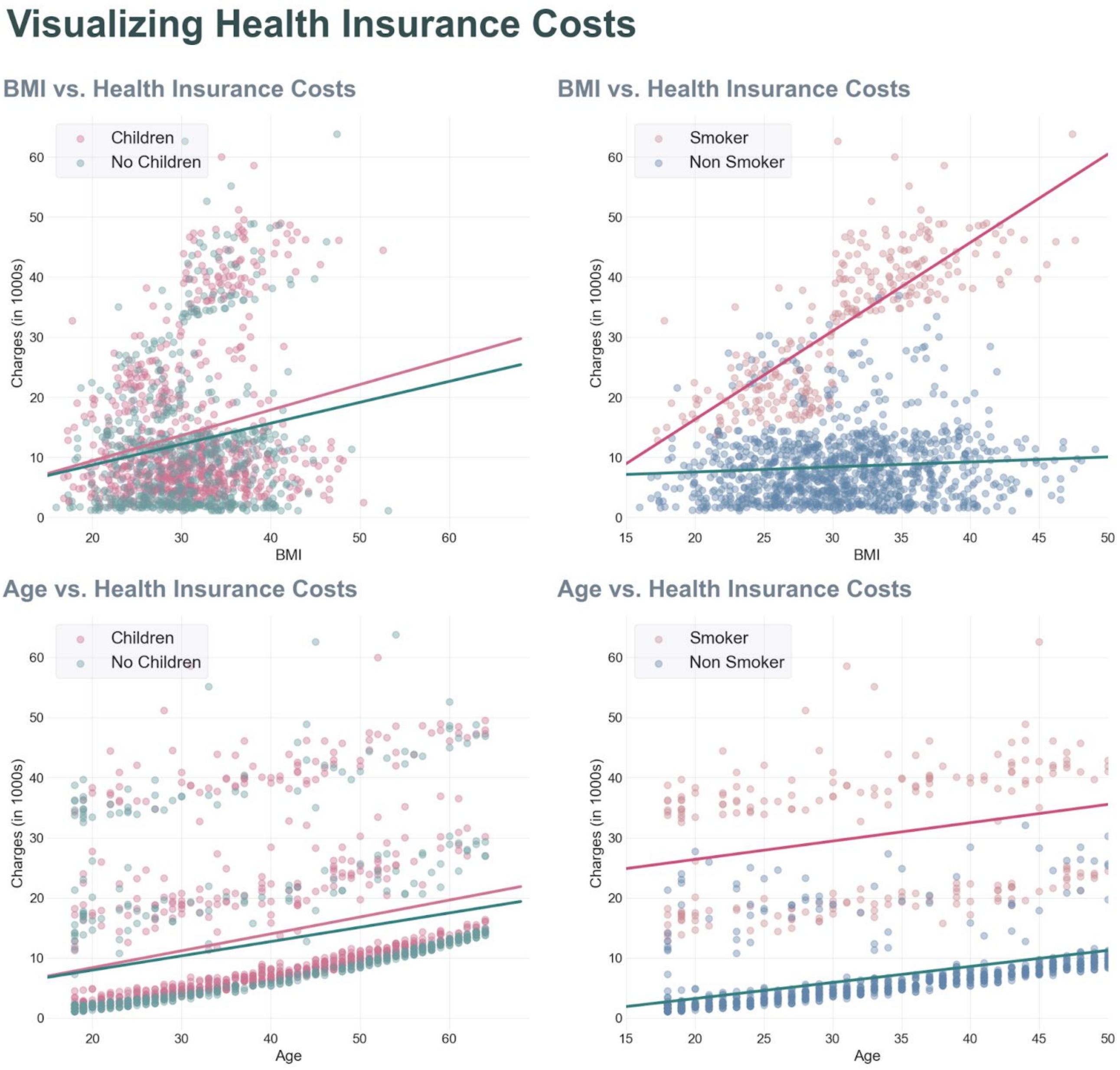
Original Plot:



Identify the Story:

This data centers around medical insurance costs and potential factors that may affect it. The plot above is a small multiple chart with scatter plots and line of best fits comparing body mass index and insurance charges on the main axes. They have also queried the data based on the number of children and plotted these subsets as the multiple charts. The plot above demonstrates that there is a positive correlation with body mass index and health insurance charges; however, this correlation appears to decrease as the number of children increases, as shown in the progression of the multiple small charts.

Improved Visualization:



Our Changes:

1. The creators of the old visualization utilized a sequential ordering of multiple charts to demonstrate how three variables may work together. Two of the three variables (bmi and number of children) are meant to explore their significance on the outcome of health insurance price.

2. Rather than using a small multiple chart, **we decided on making use of color to represent a third variable**. The first scatterplot in our redesign is a representation of the same three variables they displayed. By our line of best fit, the audience could readily see the correlation between bmi and insurance charges for the subset of those with children vs. without children.
3. We used the **principle of similarity** by making the **line of best fit the same color as the scatter plot** points with the same category.
4. We also reduced the clutter **by muting the gridlines, left adjusting our subtitles** and dividing the **truncating the y-axis** by dividing the insurance charge by 1000.
5. For our revised visualization, we decided to show how three variables intermingle by utilizing colors. The audience will now be able to quickly see the distinction between categories.

Story from the Improved Visualization:

1. Our first plot in the visualization reveals the same variables as the first design with a different story; **in our plot, the addition of children increases the correlation between bmi and health insurance charges**. We can see that the line of best fit for those with children has a larger slope than those without children. This is interesting because as we saw in the original plot, the addition of children decreases this correlation. However, as a binary subset of those with children and without children, the correlation of bmi and health charges is larger for those with children on average.
2. There was a very interesting trend that's evident in the second plot comparing bmi and health insurance costs**.** **The correlation between bmi and health insurance** costs appears to be very **strong for the subset of those who smoke.** We believe this is due to an **underlying confounding correlation** between smoking and bmi and smoking and health insurance costs.
3. For both bottom plots, **there is an evident tier of health insurance** seen by plotting age against charges. The amount that people are charged in insurance appears to be clustered into three different charge brackets demonstrated consistently over age. There is also a positive correlation between age and health insurance charge.

4. Another interesting observation by adding in that third variable of children is that **in the lower tier** charge bracket, **those with no children are consistently paying less** than those with children. However, this same finding isn’t strongly visible in the other tiers of healthcare insurance charges.
5. For the last visualization, we changed the third variable to smoking and nonsmoking. We can clearly see **a definite differentiation in the prices between those who smoke and those who do not.** At large, nonsmokers belong to the lowest tier of healthcare insurance charges while smokers are distributed across the other two tiered brackets.