وزارة الاتصالات وتقنية المعلومات
MINISTRY OF COMMUNICATIONS AND INFORMATION TECHNOLOGY

الأكاديمية السعودية الرقمية
SAUDI DIGITAL ACADEMY

CODING DOJO ACADEMY

Google Analytics

Spark Weekly Project

Data Ninjas

5/12/2022

Our Team

Renad Alahmadi

Norah Alamri

Hajer Alhoqail

Zainab Alhejji

Nouf Almuwaisheer

Anfal Almohaimeed

Outlines:

- Introduction
- Objectives
- Dataset Overview
- Pre-Processing
- Visualization Tips
- EDA
- Spark ML

Introduction

Tableau Challenges Power BI Challenges Track & Review Contributors

Workout Wednesday

A weekly challenge to re-create data-driven visualizations.

- It is a website that contains a weekly challenge to re-create data-driven visualizations.
- Challenges are released on Wednesdays.
- Solution walk-throughs are shared on Saturdays. They are at the bottom of the challenge.

https://www.workout-wednesday.com/

Objectives:

01 Check the countries with the highest number of visits to the website.

02 Observe the traffic of WOW challenges through the weekdays.

Dataset Overview

Dataset Overview:

- **Number of Columns :** 10 Columns
- **Number of Rows :** 33,865 Rows

- **Pre-Processing :** 14 Columns

Dataset Columns:

Column	Description
Country	The countries analyzed.
Date	Display the date.
Source	The Source in Google Analytics is where the website's traffic comes from. The traffic come from people visiting The site from search engines, or from a social media site or some another website. When it doesn't come from a website, or there is no data on the original website, the source is known as Direct.
Exits	The number of exits from the website.

Dataset Columns:

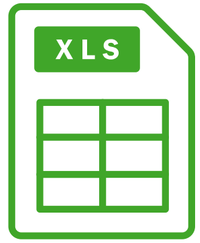
Column	Description
Medium	Medium is how the website's traffic arrived at the site. There are some core categories within Medium: - Organic Traffic (non-paid traffic from search engines) - Referral (a link from another website) - None (direct traffic)
Pageviews	Display the total number of times any pages were visited.
Session	The total number of sessions.

Dataset Columns:

Column	Description
Time on Page	Time (in seconds) users spent on a particular page.
Unique Pageviews	How many users visited a specific page.
Bounce	The total number of single page (or single interaction hit) sessions for the website.
Session Duration	Total duration (in seconds) of users' sessions.

Data Pre-Processing

Excel Pre-Processing



Dataset Pre-Processing:

1- Split the ‘Date’ column into three columns that are ‘Day’, ‘Month’ and 'Year'.

Date
12/29/2019
12/30/2019
12/30/2019
12/31/2019
12/31/2019
1/3/2020
1/7/2020
1/8/2020

Day	Month	Year
Sunday	12	2019
Monday	12	2019
Monday	12	2019

2- Split the ‘Source/Medium’ column into two columns ‘Source’ and ‘Medium’.

Source / Medium
(direct) / (none)

Source	Medium
(direct)	(none)

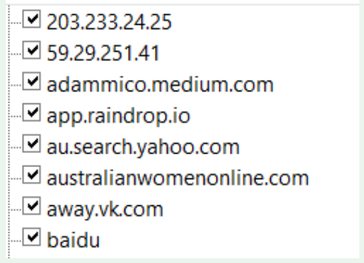
Dataset Pre-Processing:

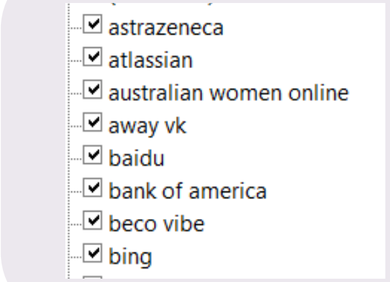
3- Removed brackets from 'Source' and 'Medium' columns to make cells more readable.

Source	Medium
(direct)	(none)

Source	Medium
direct	none

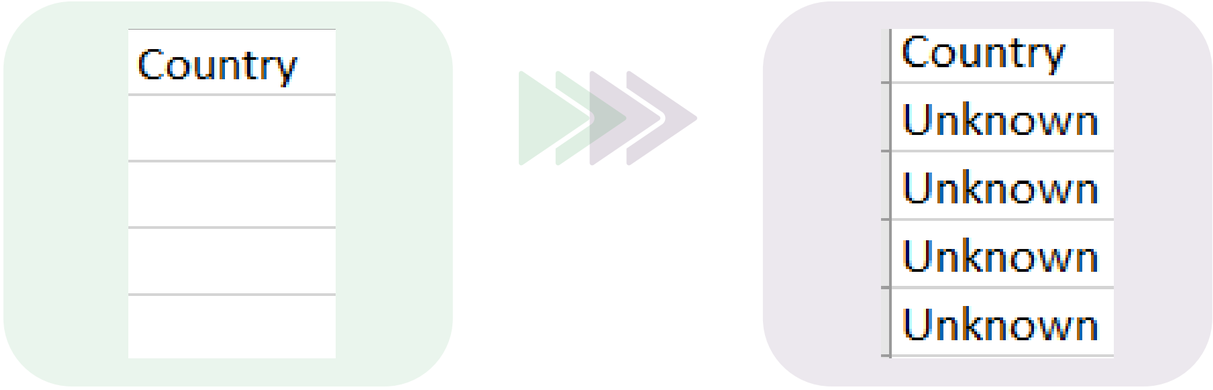
4- In 'Source' column, some sources were using IP address instead of the name, some were using URL. We cleaned them and kept only the name.





Dataset Pre-Processing:

5- In 'Country' column, there were null values, we replaced them with 'unknown'.



Python Pre-Processing

Dataset Pre-Processing:

1- The dataset is from 2019 - 2021, however, we have only 5 records that are in 2019 so we decided on dropping them to keep our analysis for only 2020 and 2021.

```
df.drop(df.index[:5], inplace=True)
```

```
# now our dataset start from 2020.
df.head(5)
```

	Country	Day	Source	Medium	Exits	Pageviews	Session Duration	Sessions	Time on Page	Unique Pageviews	Month	Year	Bounces
5	Unknown	Friday	direct	none	2	6	282	2	282	5	1	2020	0
6	Unknown	Tuesday	direct	none	1	1	0	1	0	1	1	2020	1
7	Unknown	Wednesday	direct	none	2	4	54	2	55	4	1	2020	1
8	Unknown	Wednesday	google	organic	1	3	35	1	35	3	1	2020	0
9	Unknown	Wednesday	twitter	referral	1	3	761	1	761	2	1	2020	0

Dataset Pre-Processing:

2- We had two columns with spaces so we need to rename them.

```
df.rename(columns = {'Source ':'Source', ' Medium': 'Medium'}, inplace = True)
```

```
# checking again
df.columns
```

```
Index(['Country', 'Day', 'Source', 'Medium', 'Exits', 'Pageviews',
       'Session Duration', 'Sessions', 'Time on Page', 'Unique Pageviews',
       'Month', 'Year', 'Bounces'],
      dtype='object')
```

Dataset Pre-Processing:

3- We want to create new columns to count the number of days people have visited the challenges on the website.

```
def Day_Counts(day):

    if day == 'Wednesday':
        return "Challenge Day Traffic"
    elif day == 'Thursday':
        return "One Day After"
    elif day == 'Friday':
        return "Two Day After"
    elif day == 'Saturday':
        return "Solution Day Traffic"
    else:
        return "After Solution Day"
```

```
df['CountOfDays'] = df['Day'].map(Day_Counts)
```

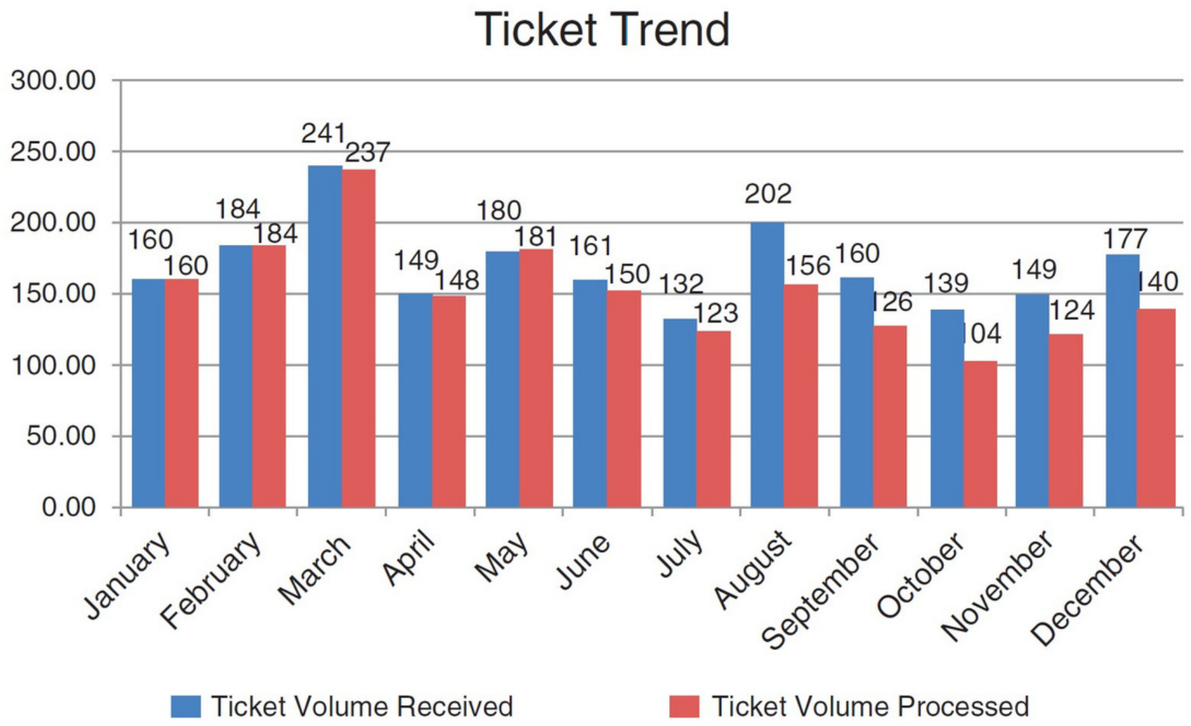
```
# checking our dataset.
df.head(5)
```

	Country	Day	Source	Medium	Exits	Pageviews	Session Duration	Sessions	TimeOnPage	Unique Pageviews	Month	Year	Bounces	CountOfDays
5	Unknown	Friday	direct	none	2	6	282	2	282	5	1	2020	0	Two Day After
6	Unknown	Tuesday	direct	none	1	1	0	1	0	1	1	2020	1	After Solution Day
7	Unknown	Wednesday	direct	none	2	4	54	2	55	4	1	2020	1	Challenge Day Traffic
8	Unknown	Wednesday	google	organic	1	3	35	1	35	3	1	2020	0	Challenge Day Traffic
9	Unknown	Wednesday	twitter	referral	1	3	761	1	761	2	1	2020	0	Challenge Day Traffic

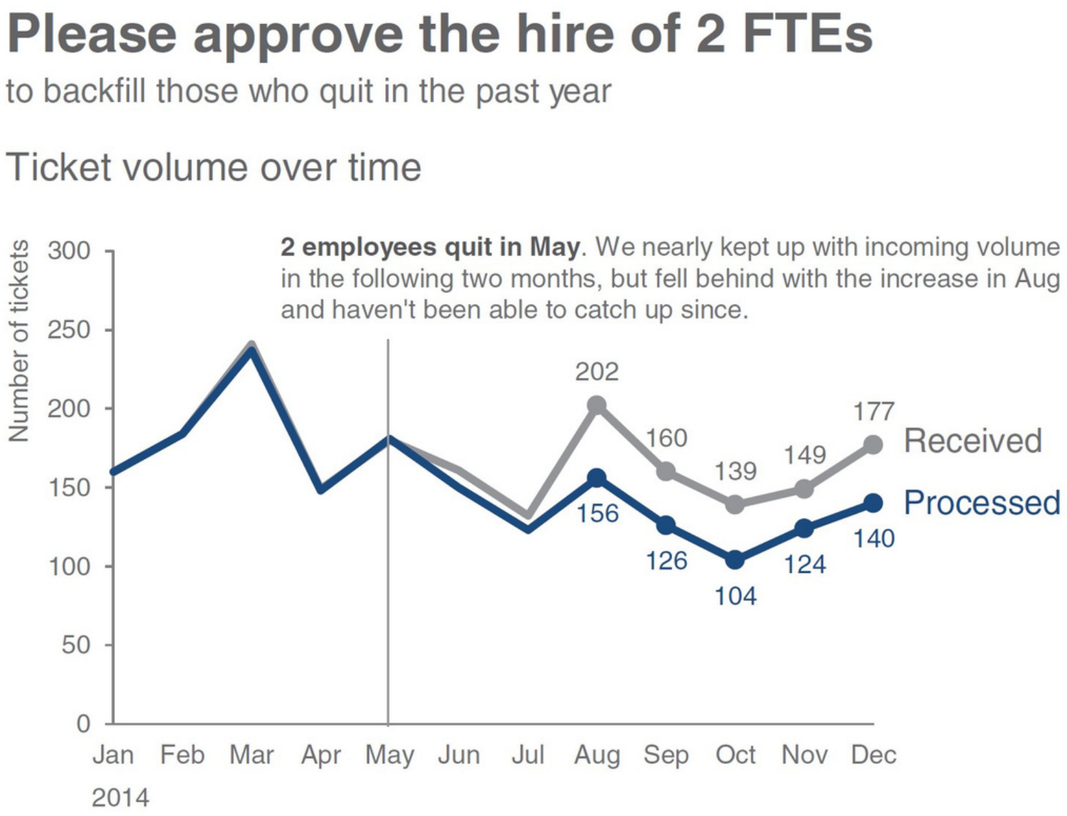
Visualization Tips

With Zainab!

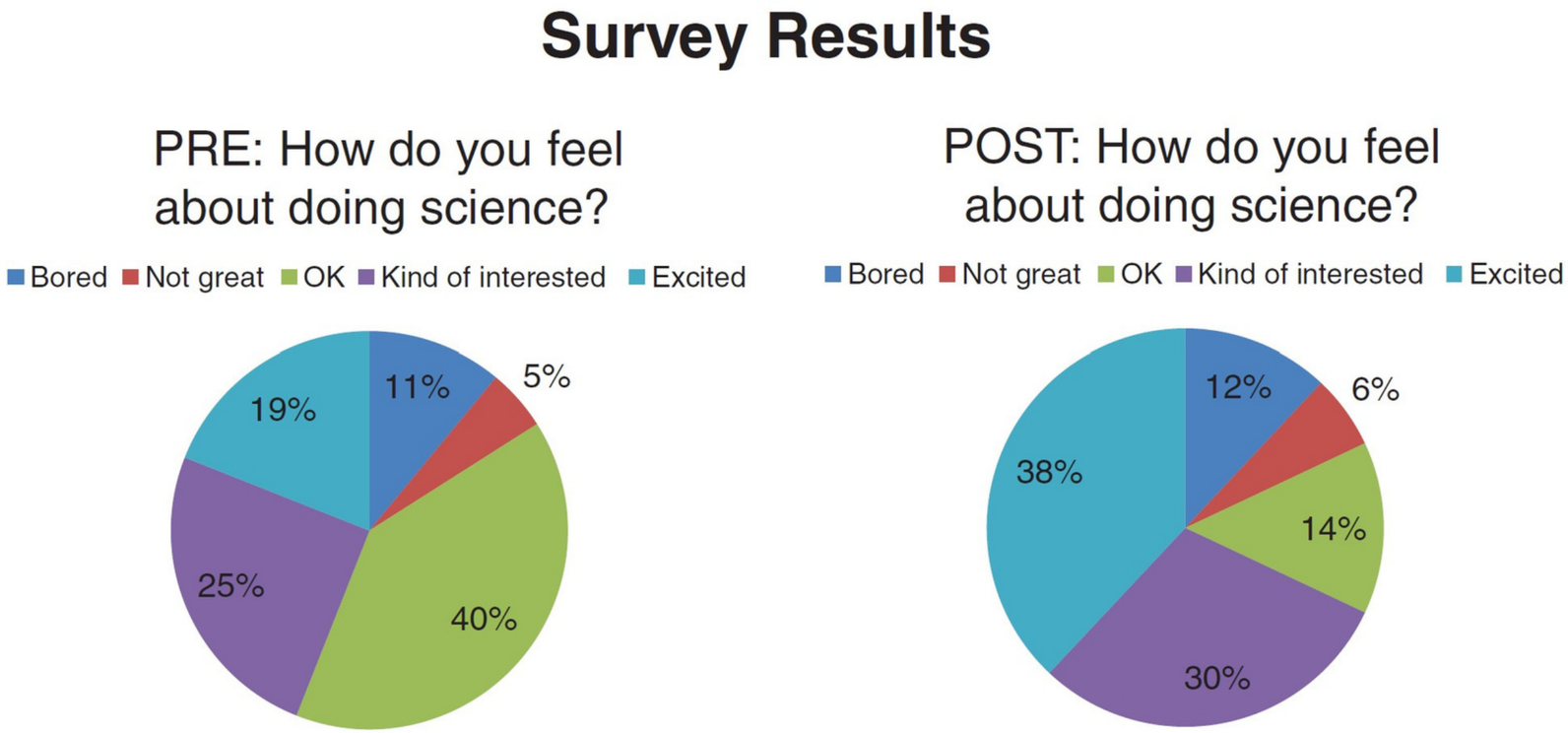
Visualization Tips: ✘



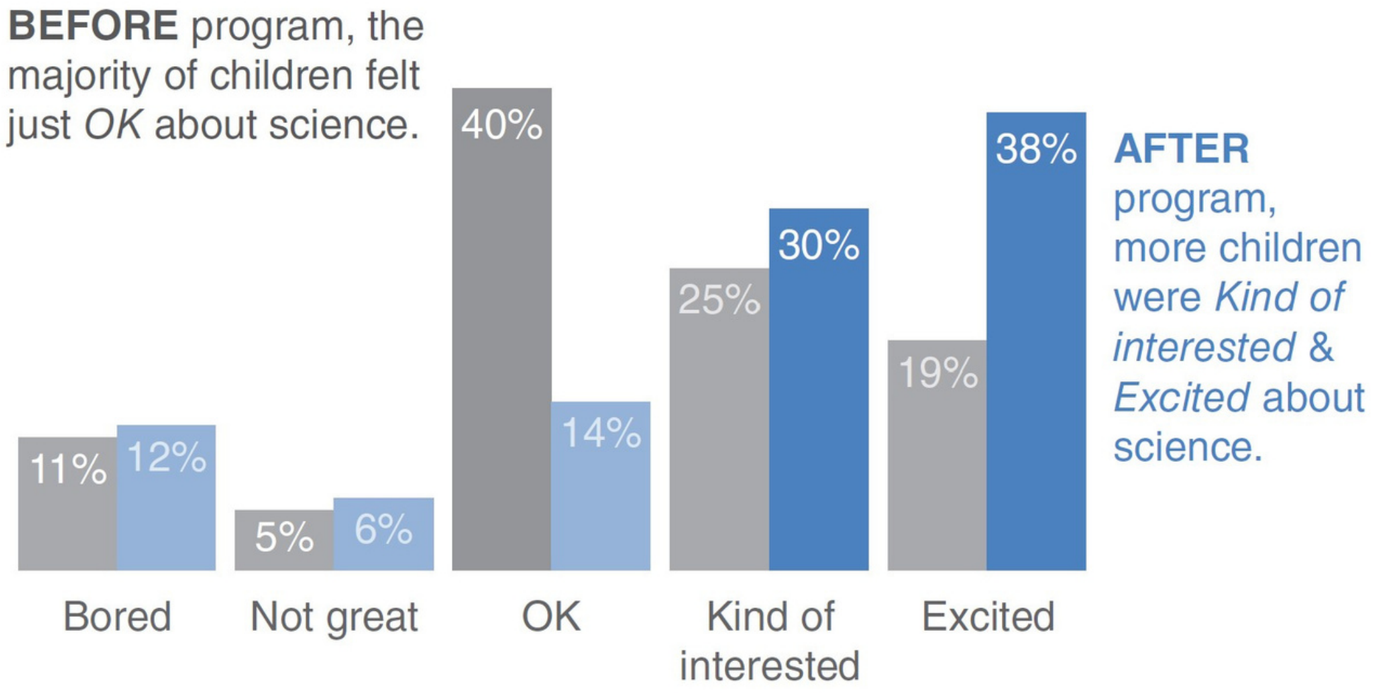
Visualization Tips: ✓



Visualization Tips: ✘

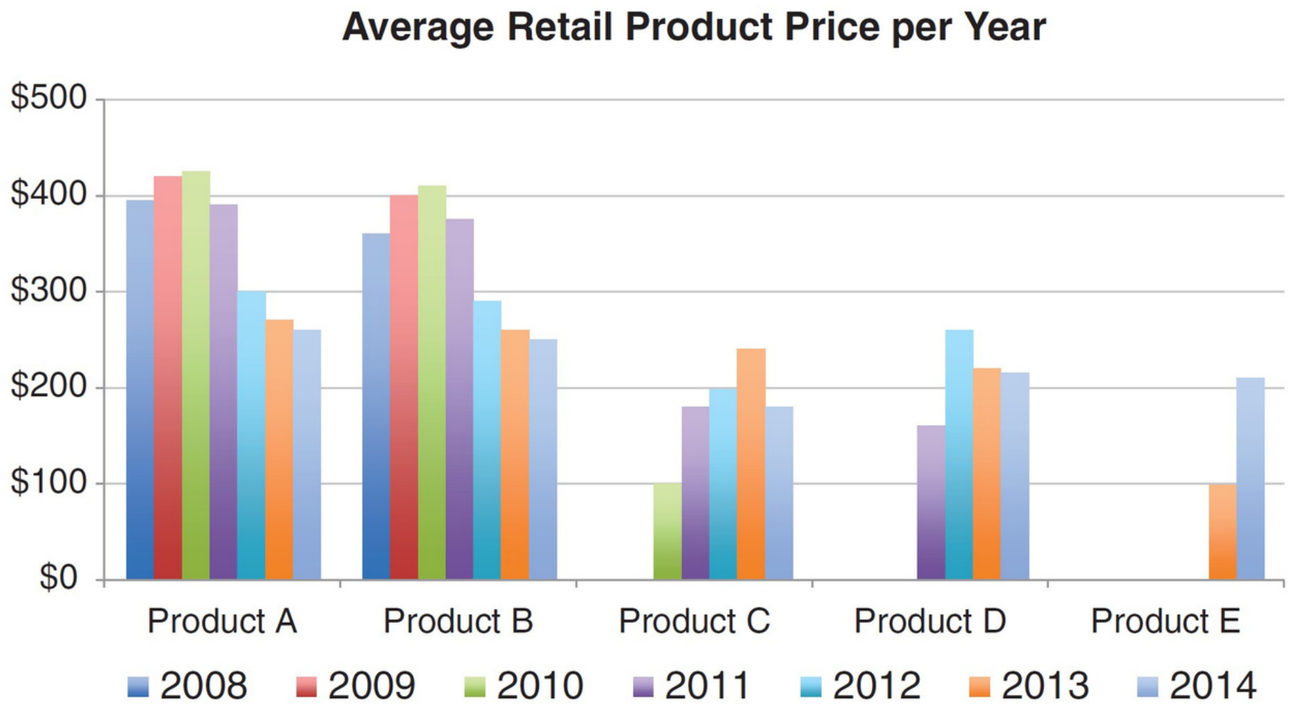


Visualization Tips: ✓

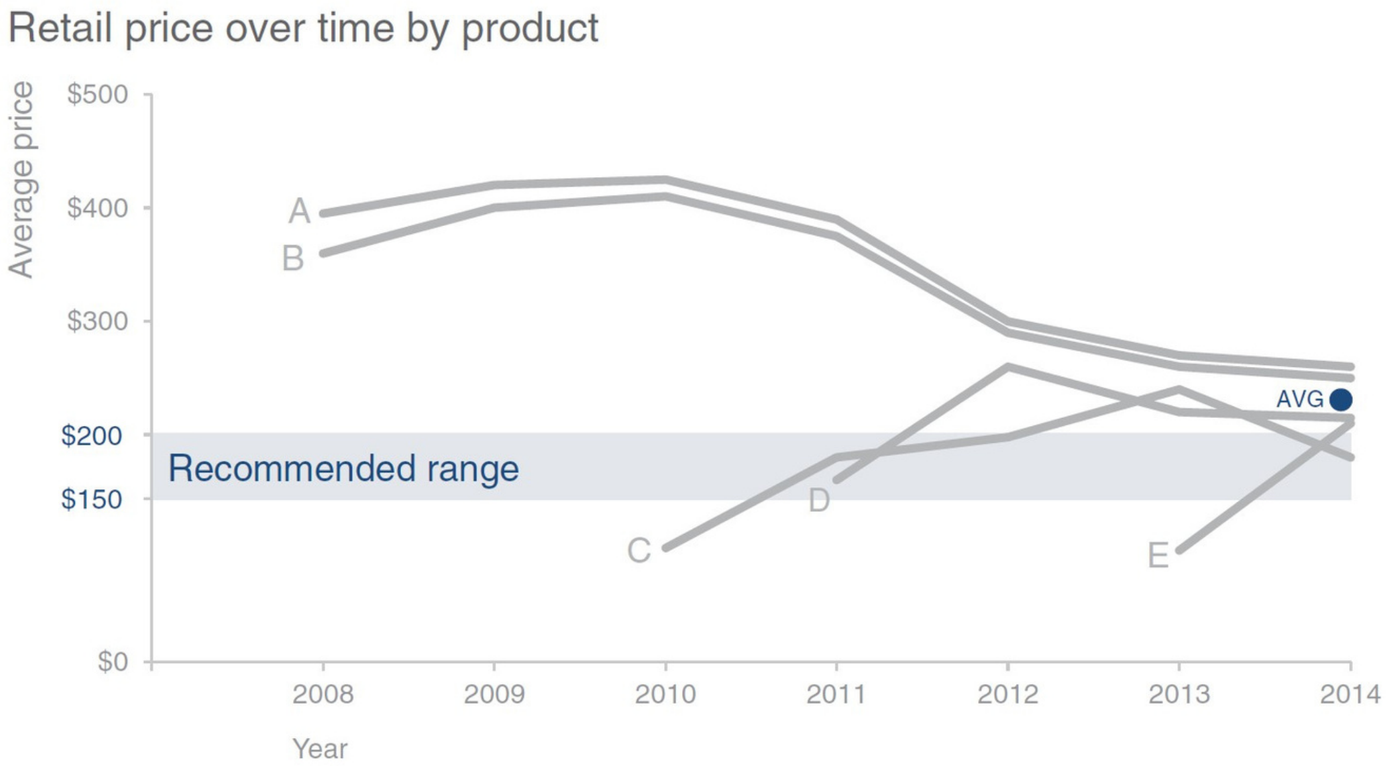


Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

Visualization Tips: ✘

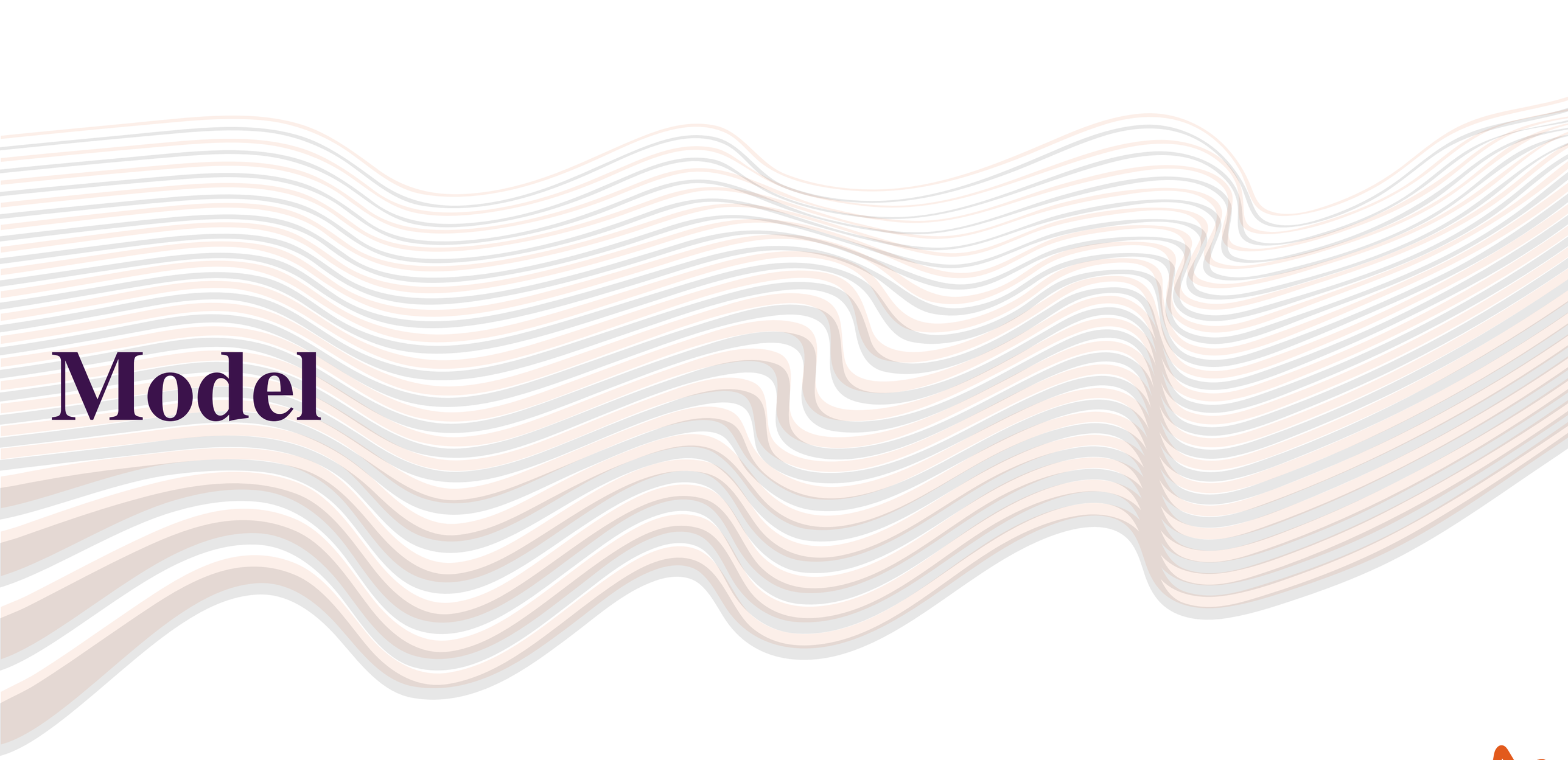


Visualization Tips: ✓



EDA

Model



Thank You!

Any Questions?

Data Ninjas

4/12/2022