Exploratory Data Analysis: Verizon Internet Service Offers

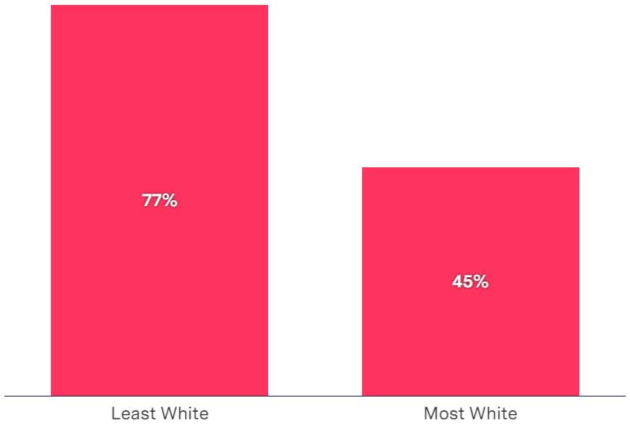
La’Tisa Ward and Lesley Frew
CS 624 Fall 2024

Problem Statement

Which is more strongly correlated with Verizon internet speed: *price* or *race of resident?*

AT&T offered least-White areas slower internet for the same price, more often

Share of households in New Orleans, La., offered slow internet, by percentage of non-Hispanic White residents



Internet download speeds of less than 25 megabits per second.

Chart: Joel Eastwood • Source: The Markup analysis of AT&T; U.S. Census Bureau

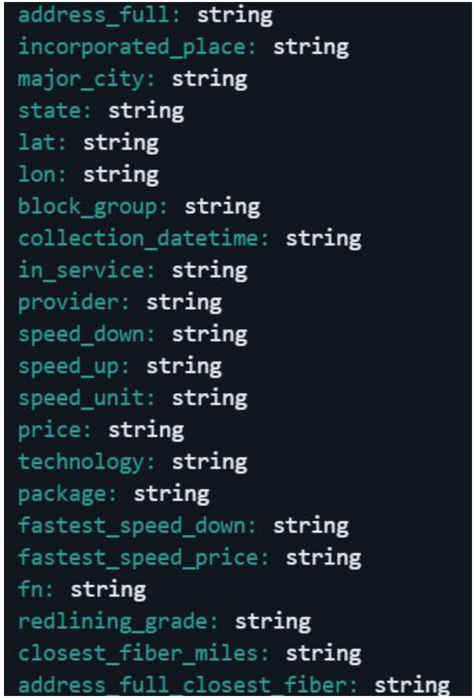
Loading with Spark

Loading: The CSV dataset loaded successfully with Spark:

```
df = spark.read.csv("/FileStore/speed_price_verizon.csv",
header=True)
```

Challenges: all of the numeric data loaded as strings

Initial observations: there are categorical variables (city, state, technology, provider, redlining_grade) and continuous variables (price, speed, race_perc_non_white).



```
address_full: string
incorporated_place: string
major_city: string
state: string
lat: string
lon: string
block_group: string
collection_datetime: string
in_service: string
provider: string
speed_down: string
speed_up: string
speed_unit: string
price: string
technology: string
package: string
fastest_speed_down: string
fastest_speed_price: string
fn: string
redlining_grade: string
closest_fiber_miles: string
address_full_closest_fiber: string
```

Sampling and Cleaning

Cleaning: We needed to drop the null rows for technology, speed_unit, and redlining_grade.

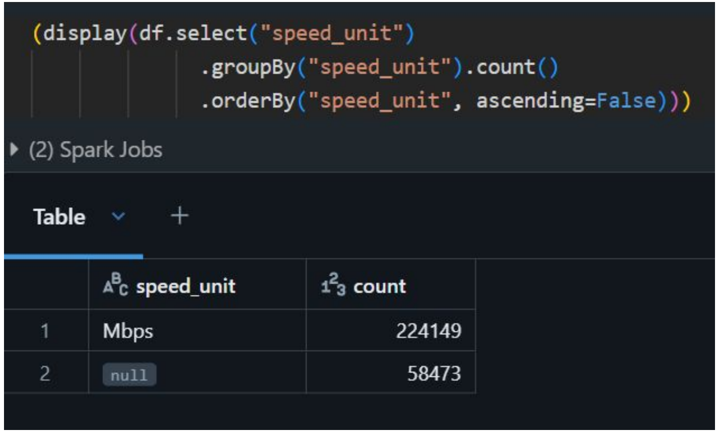
```
df3 = df.filter(df.redlining_grade.isNotNull())
```

282,622 rows ➔ 213,091 rows

Deriving: We added 2 columns: cost_per_mbps and speed_desc

```
df3 = df3.withColumn("cost_per_mbps", (col("price")
/ col("speed_down")))
```

```
df3 = df3.withColumn("speed_desc",
       when((col("speed_down") < 25), "slow")
      .when((col("speed_down") < 100), "medium")
      .when((col("speed_down") < 200), "fast")
      .otherwise("blazing"))
```



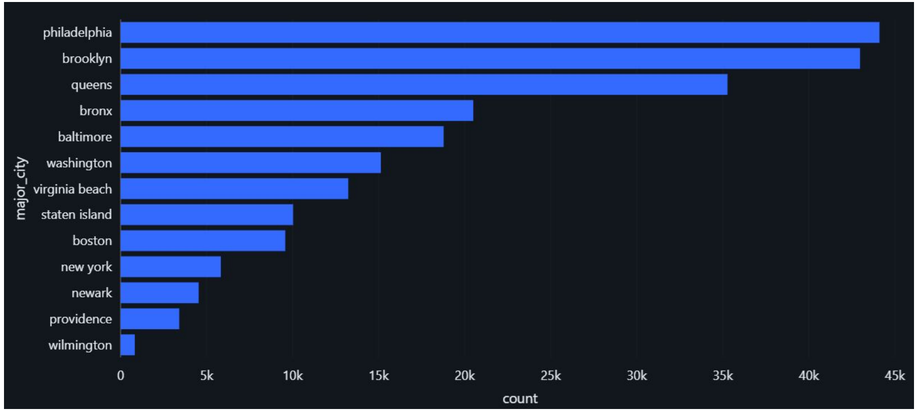
```
(display(df.select("speed_unit")
                 .groupBy("speed_unit").count()
                 .orderBy("speed_unit", ascending=False)))
```

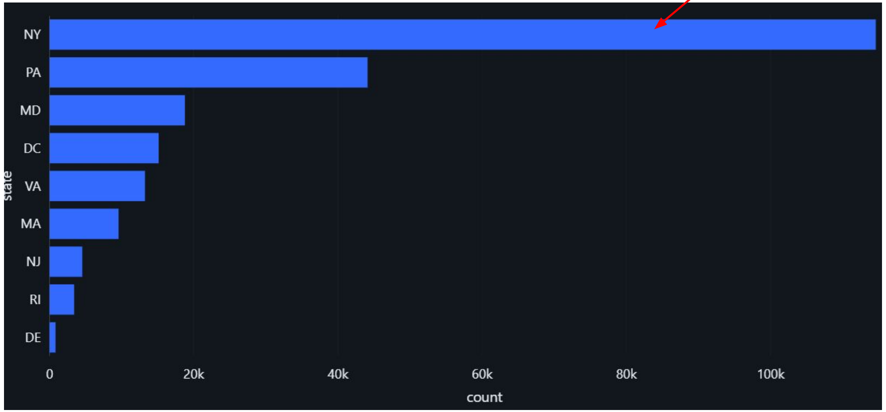
	speed_unit	count
1	Mbps	224149
2	null	58473

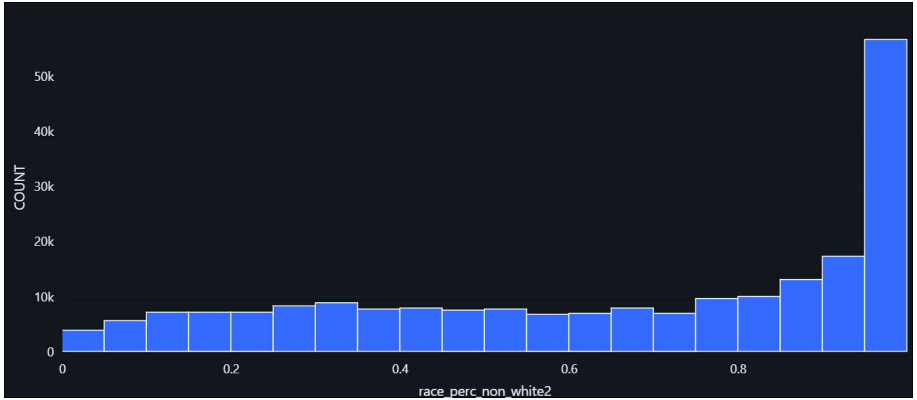
Sampling: we did not sample.

Univariate Analysis

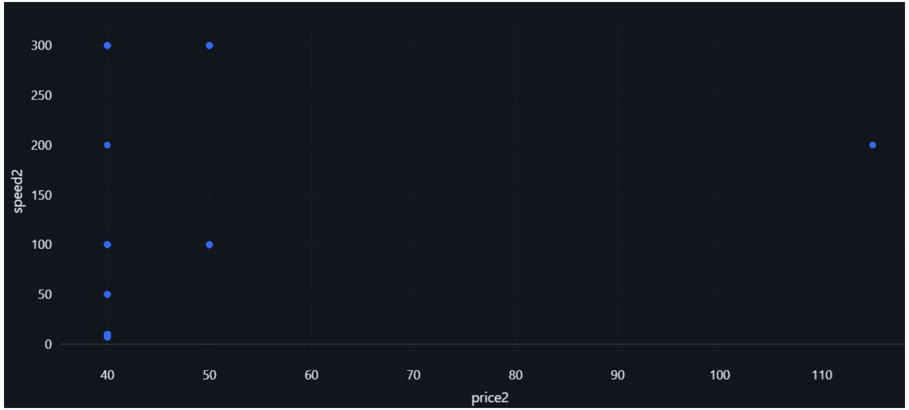
Data skew

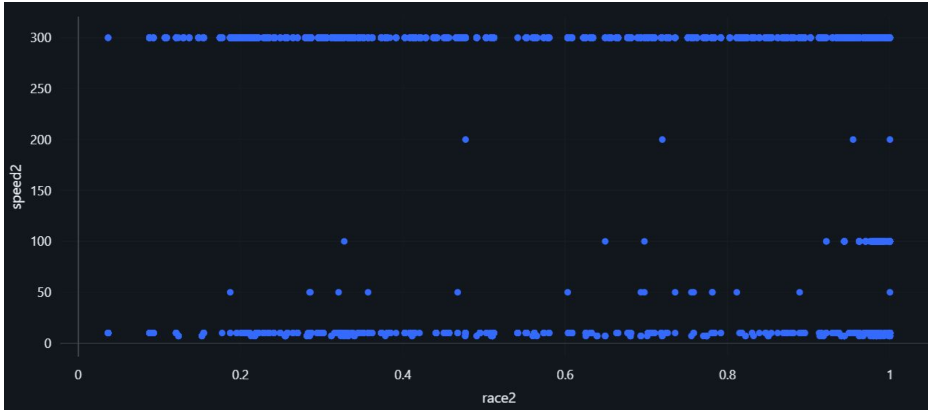






Bivariate Analysis

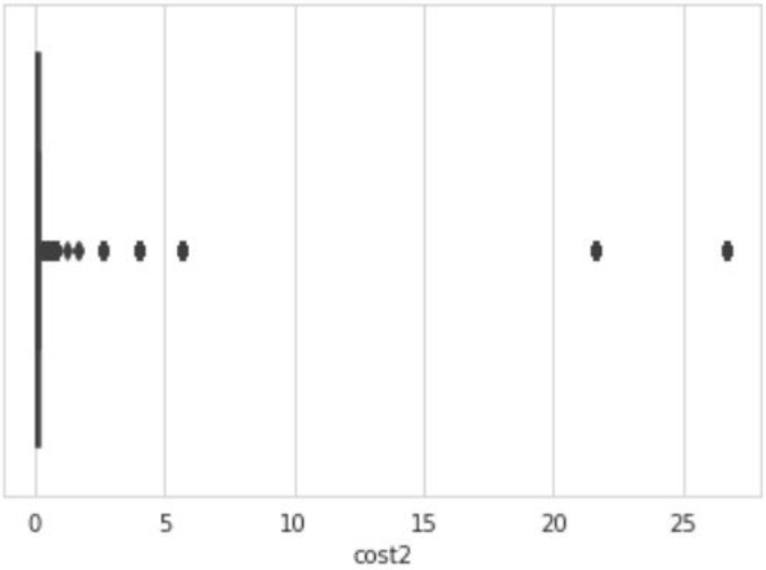




Price and speed are not correlated.

There are definitely 2 different classes of offers (fast and slow), but it’s not correlated with race_perc_non_white.

Outlier Detection Visualization

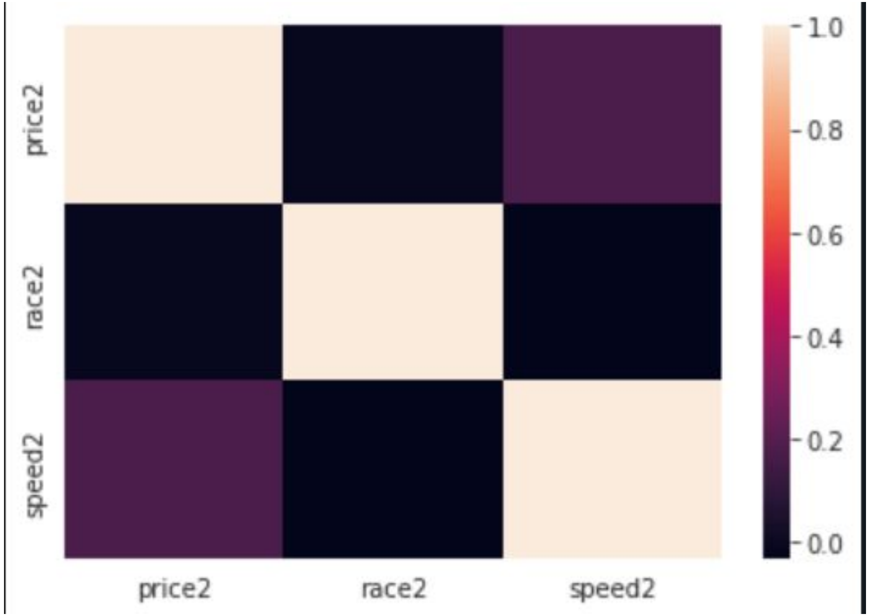


Observations:

This is a plot of cost_per_mbps. Most costs are good values (less than 1) but there are some outliers that are a poor value.

Multivariate Analysis – Correlation Maps

Observations: Race is not correlated with price or speed. Price and speed are barely correlated with a very weak relationship.


price2
race2
speed2
price2
race2
speed2