Note: Output designed for html

Tristan Macy

3/22/2021

## Introduction

This exploration revolves around the [Transit Cost Dataset](https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-01-05/transit_cost.csv) provided by the [Transit Cost](https://transitcosts.com/) project and posted on the [Tidy Tuesday](https://github.com/rfordatascience/tidytuesday) project. Given the growing concerns with climate change, public transportiation has been presented as a means to reduce city-wide carbon emissions. Given increased interest in U.S infastructure due to the recent American Jobs plan, it is important to analyze the most cost effective ways to build transit lines. The goal of the exploration is to find what barriers related to cost are present in constructing these projects within the 50+ countries present within the dataset, mainly through the **cost per km** variable. This would provide insight into which countries have found the most success in minimizing construction costs and draw conclusions to how they compare to the United States. Questions this analysis will seek to answer include:

* Which countries hold the lowest/highest **cost per km** in transit construction?
* Do higher **stations** lead to higher **cost per km** ?
* Do higher **tunnel percentages** lead to higher **cost per km** ?
* Do certain **continents** have higher **cost per km**?
* How does the U.S **cost per km** relate to other countries present within the dataset.

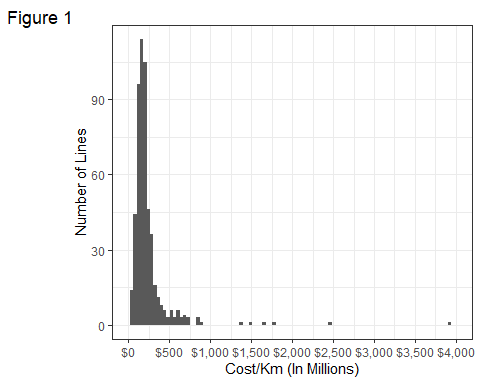
### The Data

Before any analysis could begin, there were some specific problems with how the data had been entered into R that needed be addressed. The first of these problems comes in the form of variables within the dataset being recognized as characters rather than doubles. Variables that required this specific changed were **real\_cost** , **start\_year** , **end\_year** and **tunnel\_per**. Aside from the variable changes, only a few of the entries had to be removed due to them missing substantial information such as **cost per km** and **country**.

The dataset had come with country abbreviations rather than the country names themselves. Due to unfamiliarity with these abbreviations, the countrycode library was used to create a new variable **country\_name** to use within the analysis. Later on within this project, countrycode was used to add another variable **continent** for analysis.

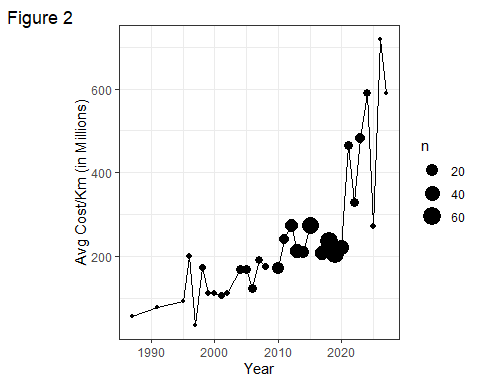
## Data Exploration

### Distribution of Transit Lines by Cost/Km



Taking a look at the histogram distribution in *Figure 1* we see that the majority of transit **lines** cost below 300 million dollars per km. There do appear to be plenty outliers above 500 range with the largest even approaching the $4000 mark. Given that the data spans across 40 years, we can expect the overall price of transit projects to gradually increase potentially explaining the higher values.

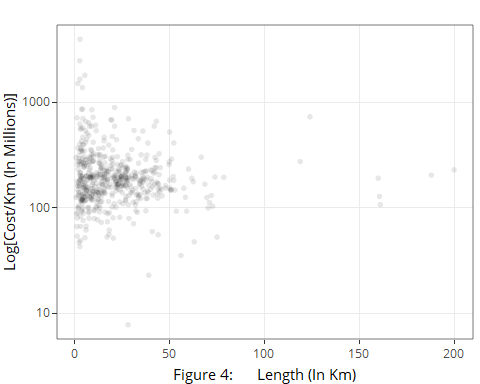
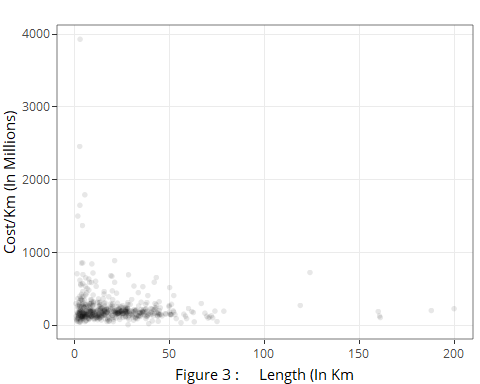
### Avg Cost/KM Across Time



As expected from *Figure 2* we can see that the average **cost per km** has steadily increased up until 2020 where we see some fluctuations on the predicted cost of future projects. These future projects may be more ambitious than the projects of the previous years. We also see that the majority of constructions took place between the years 2010-2020.

### Does Length Affect the Cost/Km?

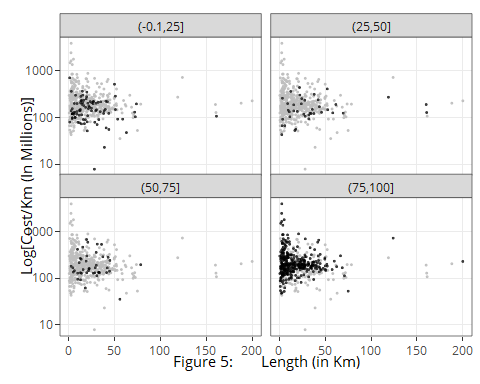
Is money saved through **cost per km** by undertaking smaller or larger scale projects?



Taking a look at *Figure 3* there doesn’t appear to be a large relationship between **Length** and **cost per km**. The highest **cost per km** lines can be identified as the U.S which are also relativley short in length. Taking the log of **cost per km** gives a better view of the data in *Figure 4*, which will be useful in identifying if other variables such as **tunnel\_per** or **stations** increase **cost per km** .

### Does Tunnel Percentage Affect the Cost/Km?

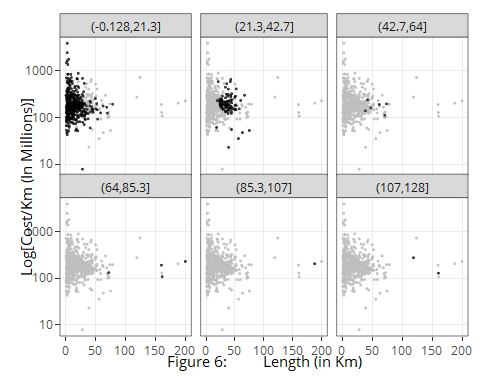
The dataset gives the percentage of each line that is tunneled through **tunnel\_per**, which likely influences the amount needed to construct it.



Faceting by 4 different bins of **tunnel\_per**, *Figure 5* reveals that majority of lines are highly tunnelled and are responsible for the most expensive of lines in terms of **cost per km**. While the difference is visible within the highly tunneled category, there does not appear to be much of a difference between the other 3 levels of **tunnel\_per**.

### Do the Amount of Stations Affect the Cost/Km?

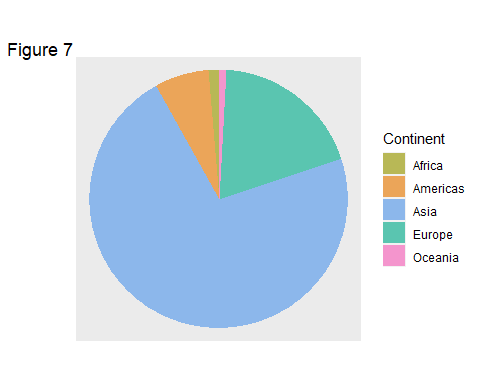
The dataset also gives us the number of transit **stations** within each line, which might prove to be a factor in determining **cost per km**.



Faceting on 6 different **station** bins in *Figure 6* reveals suprising results. While the amount of **stations** appear to increase with the **length** of a transit line, the actual **cost per km** appears to be consistent with even the highest vales occupying the lowest bin of **station**. **stations** thus, appear to not be as significant on **cost per km** as predicted.

## Continental Analysis

Now that the general analyis of the entire data is concluded, we can now analyze differences based on the locations of all the lines. Since the data spans 50+ countries, we will introduce a **continent** grouping to make draw conclusions on any differences and similiarities.



Continent

N

Proportion

Africa

7

0.0130354

Americas

36

0.0670391

Asia

387

0.7206704

Europe

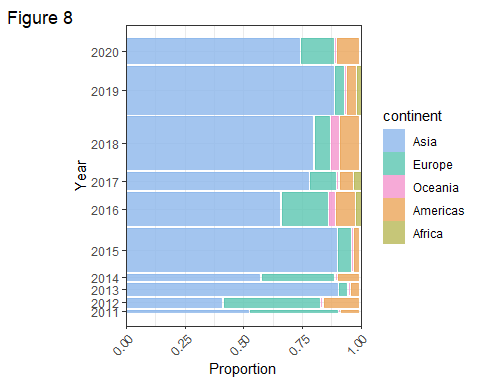
102

0.1899441

Oceania

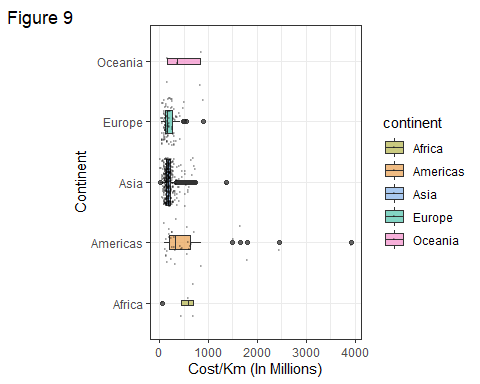
5

0.0093110



From *Figure 3* we could see the overall proportion of new railways being built through the **years** 1987 - 2027 were predominatley in **Asia** with 72.1% followed by **Europe** with 19.0% , **Americas** with 6.7%, **Africa** with 1.3% and **Oceania** with 0.9%. Taking a closer look between years 2010-2020 within, *Figure 8* we see a similiar trend within the last 10 years with construction being dominated by the combination of **Asia** and **Europe**.

### Continental Boxplots



*Figure 9* reveals that the **Americas** appear to have a larger median and range than that of **Asia** and **Europe**. The **Americas** also boast the largest outliers that we had found earlier to be due to the United States. Despite the large amount of projects undertaken by **Asia** and **Europe**, they consistently maintain a lower **cost per km** in their transit construction. **Oceania** and **Africa** maintain higher medians than the other 3, but have too little entries to draw meaningful conclusions. A closer look in the **Americas** may reveal that the United States may be responsible for the increased **Cost per Km**.

## Closer Look at the United States

### How Does the U.S Stack Up Against Other Countries?

The top 10 countries with constructed lines in the dataset were taken and their Average **cost\_km\_millions** computed. The rest of the countries were grouped into a other category.

Country Name

N

Avg Cost/Km (Millions)

China

253

184.39462

Other

141

NA

India

29

186.56489

Turkey

20

107.96300

Spain

15

97.16001

France

15

183.38034

Japan

15

241.25823

Germany

13

251.64929

United States

13

1211.46913

Taiwan

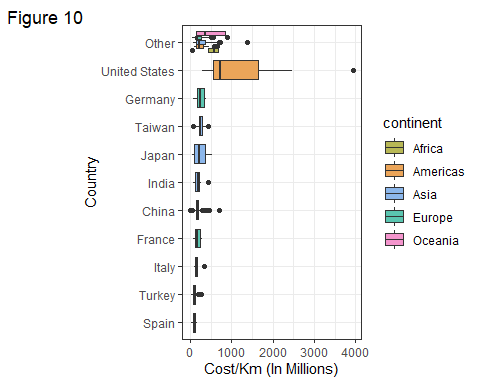
12

246.93726

Italy

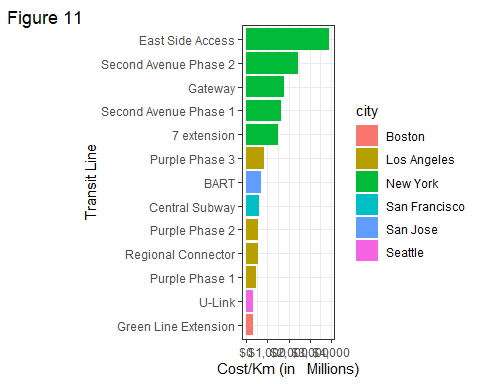
11

159.88190



In *Figure 10* we can clearly see that the U.S is much higher **cost per km** than that of any other countries. Even among more devoloped countries such as Germany and Japan, the U.S appears to still spend much more per km. It also becomes apparent the outliers within the Americas in *Figure 9* were a result of the U.S and many of the outlier values are not considered outliers within the U.S only boxplot. The U.S average of 1211.47 is also the highest within the whole dataset.

### What Transit Lines in the U.S Have the Highest Cost/KM?



Within *Figure 11*, we can see that the most expensive transit lines are in New York. When compared to New York, Los Angeles lines maintain at least half the **cost per km** of New York while only having 1 less line in total constructed. If New York were to be taken out, we would see a closer resemblance to the other boxplots within *Figure 10*. New York being the only city within the U.S with a drastic difference may prove hopeful for the future of transit projects within the U.S.

## Conclusion

Transit lines since, 1987 have largely been undertaken by Asia and Europe who likely value public transportation more than the motorized nature of the U.S. Generally the number of **stations** does not appear to have much of an impact on **cost per km** meaning costs are largely based on other aspects of construction. **Tunnel per** was revealed to include the most expensive transit lines including the most expensive lines in New York. While this is true for high **Tunnel per**, there appears to be little impact on **cost per km** for transit lines that arn’t completely (100%) tunneled.

In regards to continent, the Europe and Asia comparison reveals that the United States’ large **cost per km** is a problem that is unique to them. Compared to other countries both developed and developing, the United States is spending much more per kilometer. Perhaps the United States’ could save money from learning from the construction methods of other countries. A closer look into the United States however, reveals that the issue in spending lies mainly in spending within New York. Future analysis would thus, focus on why New York is much more expensive when compared to the rest of the U.S. Were the costs unavoidable due to specific problems, or should all future construction projects model other cities within the U.S that resemble the cost of the rest of the world?