**DSC530-T301 Data Exploration and Analysis**

**Final Analysis Summary**

**Frank Neugebauer**

Statistical/Hypothetical Question

Which factor(s) have the greatest impact on performance in my cycling data?

Outcome of your EDA

The exploration went very well in that:

* I learned a great deal about the underlying data details
* I learned how to use the Stava API (this was hard, even for an experienced programmer)
* The data is not nearly as well defined as I thought (from a statistical perspective)
* I got to explore many ways of observing and wrangling the data; while time didn’t allow for a thorough examination and transformation, it’s apparent to me where this needs to go (e.g., time-series analysis, more data transformations)
* I got to see the work done in R replicated in Python to better understand when to use each

What do you feel was missed during the analysis?

More data transformations are really required because some of the data does not follow a normal distribution as well as I’d have liked. Additionally, outliers and missing data could have been managed more thoroughly. Finally, binning should have been used for very specific combinations of data – e.g., all virtual rides could have shown more interesting results because all the data is very controlled.

Statistically, more time is required to deeply understand what the statistics mean across many predictors in a multiple regression model. While understood in isolation, this aspect is more difficult (for me) when many predictors are involved.

One final note – I believe there are aspects of performance that are difficult to measure. For example, mental state, rider weight, and other activity types before and after top performances. It’s actually hard to accurately understand top performances (except for races, but I’ve crashed as often as I’ve finished, so that may not work for me – I never said I was a good cyclist).

Were there any variables you felt could have helped in the analysis?

No. I picked all the important variables available to me. However, I could have cleansed them more – for example, some data was for non-training rides and could probably have been eliminated.

Were there any assumptions made you felt were incorrect?

Yes – that the data was thorough and complete. The data is actually incomplete not only because of missing values but also because of missing attributes (e.g., rider weight).

What challenges did you face, what did you not fully understand?

How to download Strava data using the API. I spent at least 40 hours figuring this out. The root problem was the overloaded use of the term ‘token’ by the Strava development team (and their subsequent failure to mention this).