In [11]:

```
# libraries
import pandas as pd
import numpy as np
```

PSTAT 100 Project plan

This is a guide to preparing your project plan. It functions both as a guide to the work you'll need to do and as a guide to preparing the deliverable. You can use it as a template to draft the plan report; if so, **please remove the text explanations and instructions in each section so that it reads as a coherent and continuous document**.

While you may find it useful initially to follow the outline given, you do not need to adhere to it exactly -- you're free to organize your submission in the way that seems most natural to you. However, please do keep the high-level sections, so that your report includes the following headers:

1. Background
2. Data description
3. Initial exporations
4. Planned work

Your report does not need to be long. It should be about 2-4 pages, and might not be much longer than this template once you replace the guiding text with your own work.

Group information

Group members:

- Tufei Cai
- William Fitzhugh

Contributions:

1. William worked on tidying the dataset, prepared the data description, create some data visualization, and wrote the background information.
2. Tufei worked on tidying the dataset, created some data visualization, finish exploratory analysis.

---

0. Background

This section should introduce your reader to the general topic you're engaging with in your project and explain any specialized knowledge that they may need to understand your dataset and why it's interesting. It doesn't need to be long, but should touch on the following points:

- Introduce the topic of your project.
- What area or areas of study are you in dialogue with for your project?
- What is your data about, broadly?
- What is the motivation for collecting the kind of data you're working with, and what sorts of things could you potentially learn?

You can look to the background sections in the homework assignments for examples. (There you can also see how to include images in your notebook.) The background sections of the homeworks are usually short and focused paragraphs intended to orient you to what you'll do in the assignment. They don't go into a lot of detail -- just enough to (hopefully) convince you that the data are interesting and explain any terminology or general information you may not know.

You may find it useful to write up the data description first, think about what the reader should know before they peek at your dataset, and then come back to the background section. I often write the background sections of your assignments last, once I have a sense of what kind of information would be most useful going into the assignment.

Soccer is the most popular game in the world. Played by billions of people, almost everyone has some knowledge of the game and their country's league. However, soccer is played differently throughout the world, with certain countries known for their flair and others known for their discipline. The goal of our project is to explore how these differences manifest themselves in data. The main data we will be working on is a collection of match results from eleven different European soccer leagues over a number of years (roughly 2010-2016). We will combine this data with information on player abilities, team play styles, and betting odds, to create insights into how these top leagues differ from one another. The infomation we create in this project could be used to aid the accuracy of a model used to predict soccer results.

---

1. Data description

This section should introduce your dataset in detail. It should reflect your having gone through the collect/acquaint/tidy stages of the lifecycle. Below I've provided you with an outline. You do not need to adhere to this strictly -- in fact, it would be more natural to divide the items among a few short paragraphs -- but you should touch on each item in a format that suits your project.

Basic information

Help your reader understand what your data is, where it came from, and how it can be used. Provide the following.

General description: provide a one- or two-sentence description of the data right at the beginning. For instance, "The data are diatom counts sampled from evenly-spaced depths in a sediment core from the gulf of California." Nothing too complicated, just something to give your reader a sense of the 'what' right off the bat.

Source: indicate where your data came from. Provide a verbal description -- who collected it as part of what project and where -- and either a citation or a hyperlink.

Collection methods: How were the data values obtained? Provide a simple description of how measurements were taken (using scientific equipment? web scraping? surveys?).

Sampling design and scope of inference: Indicate the relevant population. If identifiable from data documentation, state the sampling frame and sampling mechanism and indicate the scope of inference. If no information is available about the sampling design, indicate this instead, and discuss the extent to which having no scope of inference is a limitation for the particular topic you're investigating.

The match data is contains the home and away teams involved in the match, the goals scored by each team, and the result information. This data comes from a kaggle project (https://www.kaggle.com/datasets/hugomathien/soccer) which aimed to collect data on European club soccer. Data was collected via web scraping, as all of this infomation is readily availbale online, but needed to be collected and stored. The relevant population is the eleven leagues for which matches are recorded, all of which are in Europe. We will focus on the biggest leagues, being England, Spain and Italy. The scope of inference is small, if anything, because this data is years old. If this were up to date data, then our finding would have some bearing on how these leagues will perform going forward, but given that this data is at least five years old, the leagues would have had enough time to change that our finding will likely have little to no bearing on European soccer today.

Data that we will pair with the match infomation is the player and team attributes, which are taken from Fifa, the most popular sports video game in the world. Fifa is a soccer video game which maticulously tracks player abilities and team styles throughout the year. While these stats are obviously not a direct representaion of real life, they provide a good idea of how teams should play on paper. This data was also web scraped, and was found on the same kaggle page.

Data semantics and structure

Units and observations: State the observational units.

Variable descriptions: Provide a table of variable descriptions. If your dataset is large and you'll only work with a subset of the total available variables, limit your attention to the variables that you'll work with. Here's a template you can work with:

Name	Variable description	Type	Units of measurement
GCLASS	Graduating class	Numeric	Calendar year

Example rows: Print a few example rows of your dataset in tidy format. Please don't include the codes you used to manipulate the raw data. Do that in a separate notebook and export the result to a .csv file -- `data.to_csv('tidy-data.csv')` -- to load directly into the cell below.

This is the data we are sure we will work with. Some of our initial exploration will highlight some of the other data we might work with.

In [5]:

```
# load tidied data and print rows
Matches = pd.read_csv('Matches')
Matches.head()
```

Out[5]:

	match_api_id	id	League	Country	season	stage	home_team_goal	away_team_goal	home_team	away_team	winner	home_points	away_p
0	492473	1	Belgium Jupiler League	Belgium	2008/2009	1	1	1	KRC Genk	Beerschot AC	Draw	1	
1	492474	1	Belgium Jupiler League	Belgium	2008/2009	1	0	0	SV Zulte-Waregem	Sporting Lokeren	Draw	1	
2	492475	1	Belgium Jupiler League	Belgium	2008/2009	1	0	3	KSV Cercle Brugge	RSC Anderlecht	RSC Anderlecht	0	
3	492476	1	Belgium Jupiler League	Belgium	2008/2009	1	5	0	KAA Gent	RAEC Mons	KAA Gent	3	
4	492477	1	Belgium Jupiler League	Belgium	2008/2009	1	1	3	FCV Dender EH	Standard de Liège	Standard de Liège	0	

In [20]:

```
# Data descriptions
Data_Descriptions = pd.read_csv('Data_Descriptions')
Data_Descriptions
```

Out[20]:

	Variable	Variable description	Type	Units of measurement
0	match_api_id	Match id	Numeric	individual game
1	id	league id	numeric	NaN
2	League	League name	String	NaN
3	Country	Country Name	String	NaN
4	season	season of match	String	soccer calender year (fall to summer)
5	stage	stage of season	Numeric	# game of season
6	home_team_goal	Goals scored by home team	Numeric	num of goals
7	away_team_goal	Goals scored by away team	Numeric	num of goals
8	home_team	Home team name	String	NaN
9	away_team	Away team name	String	NaN
10	winner	Winning team name	String	NaN
11	home_points	Points earned by home team	Numeric	win, draw or loss
12	away_points	Points earned by away team	Numeric	win, draw or loss

2. Initial explorations

At this stage, you may spend most of your effort on the computing side tidying up the data. You're not expected to complete a thorough exploratory analysis, and if your dataset was especially messy to start with, you may not even begin your exploratory analysis by the time you prepare this report. You have the option to leave exploration for the next stage of work and simply report basic properties of the dataset, but you should at minimum address the items in the 'basic properties' section below.

Basic properties of the dataset

Help the reader get acquainted with your dataset on a simple level by identifying characteristics of the dataset and variable summaries. Some amount of code is fine here, but try to use code cells sparingly.

Dimensions: 29580 X 13

Missing values: There should be no missing values.

Variable summaries:

Variable Name : home_team_goal

- Max: 10
- Min: 0
- Mean: 1.5445937103044767
- Variance: 1.682619454408162
- Stand Deviation: 1.2971582225804845

Variable Name: away_team_goal

- Max: 9
- Min: 0
- Mean: 1.1609376804341969
- Variance: 1.30441602733493
- Stand Deviation: 1.1421103393871057

Variable Name: home_points

- Max: 3
- Min: 0
- Mean: 1.6300473459332538
- Variance: 1.7253593441287172
- Stand Deviation: 1.3135293465045679

Variable Outline

- League - The name of the League that we focus on
- Country - The place where the League located
- Season - Season time of the League
- Stage - The stage where the data collected
- home_team_goal - Total goal for the home team
- away_team_goal - Total goal for the away team
- home_team - The name of home team

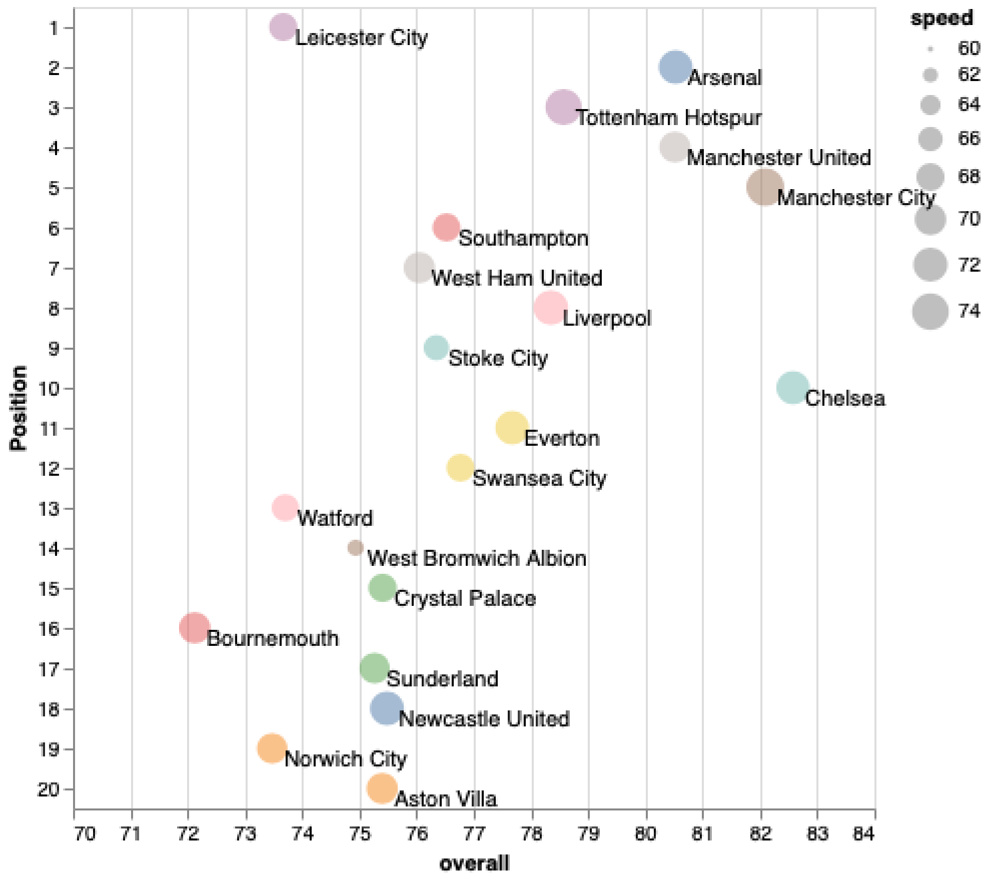
- away_team - The name of away team
- winner - The name of winning team
- home_points - Home team status(Teams receive three points for a win, one point for a tie, and zero points for a loss.)
- away_points - Away team status(Teams receive three points for a win, one point for a tie, and zero points for a loss.)

Exploratory analysis

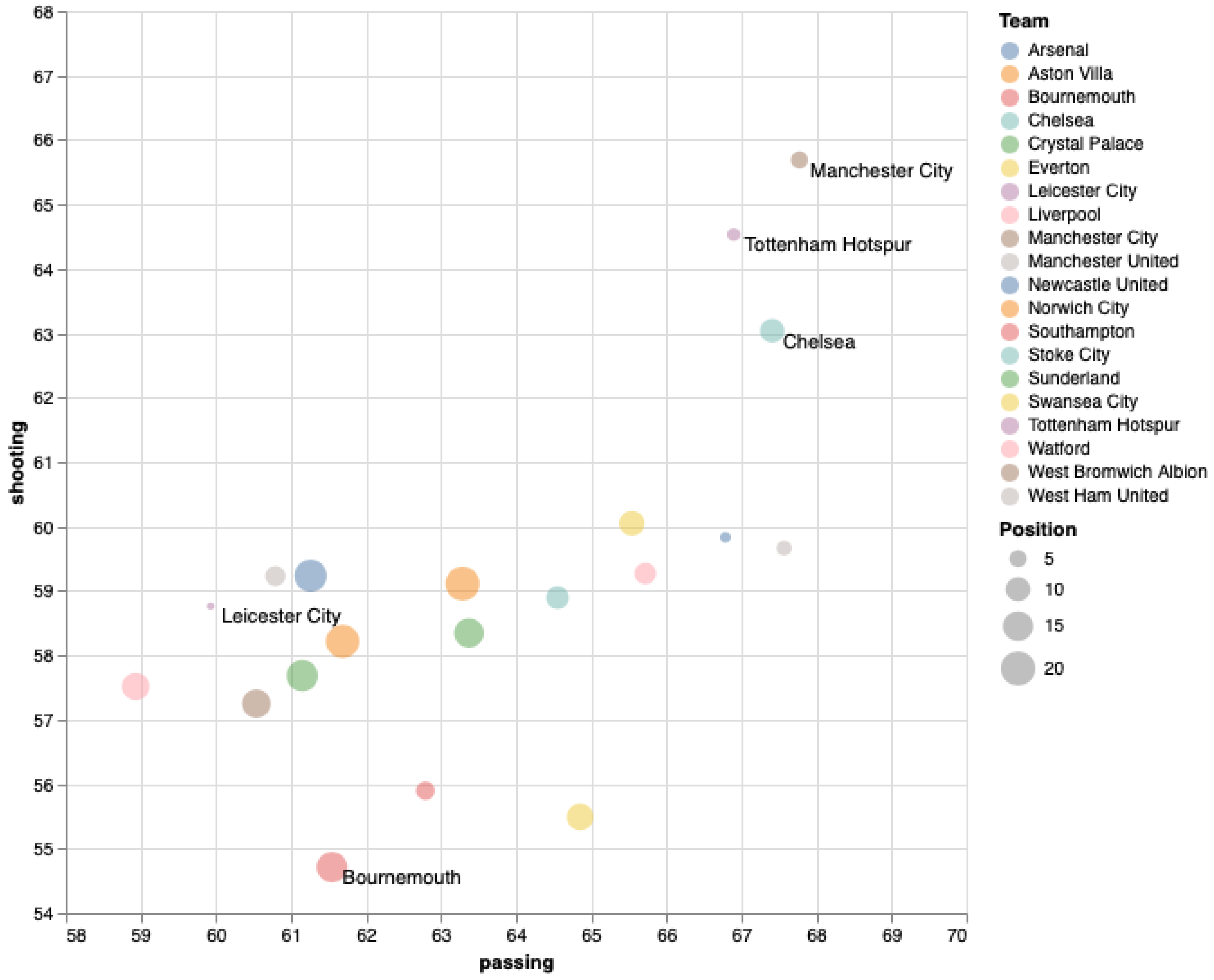
If you were lucky and your dataset was neat, you should aim to include a few exploratory plots or tables here -- they don't need to be polished at this stage, but you should select plots that are informative (rather than including all plots you may have looked at).

If you do include exploratory graphics or tables, please explain in a sentence or two what each one shows. Try to include a minimum of code. Consider saving your plots as images and inputting images into markdown cells instead of generating them anew via code cells.

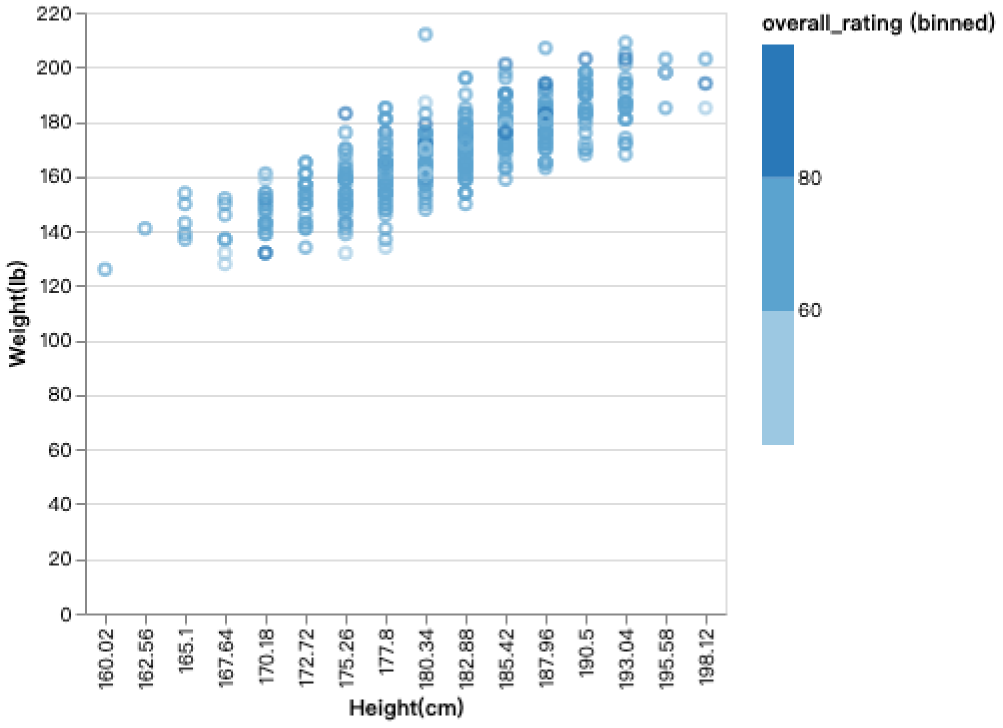
Exploratory Analysis:



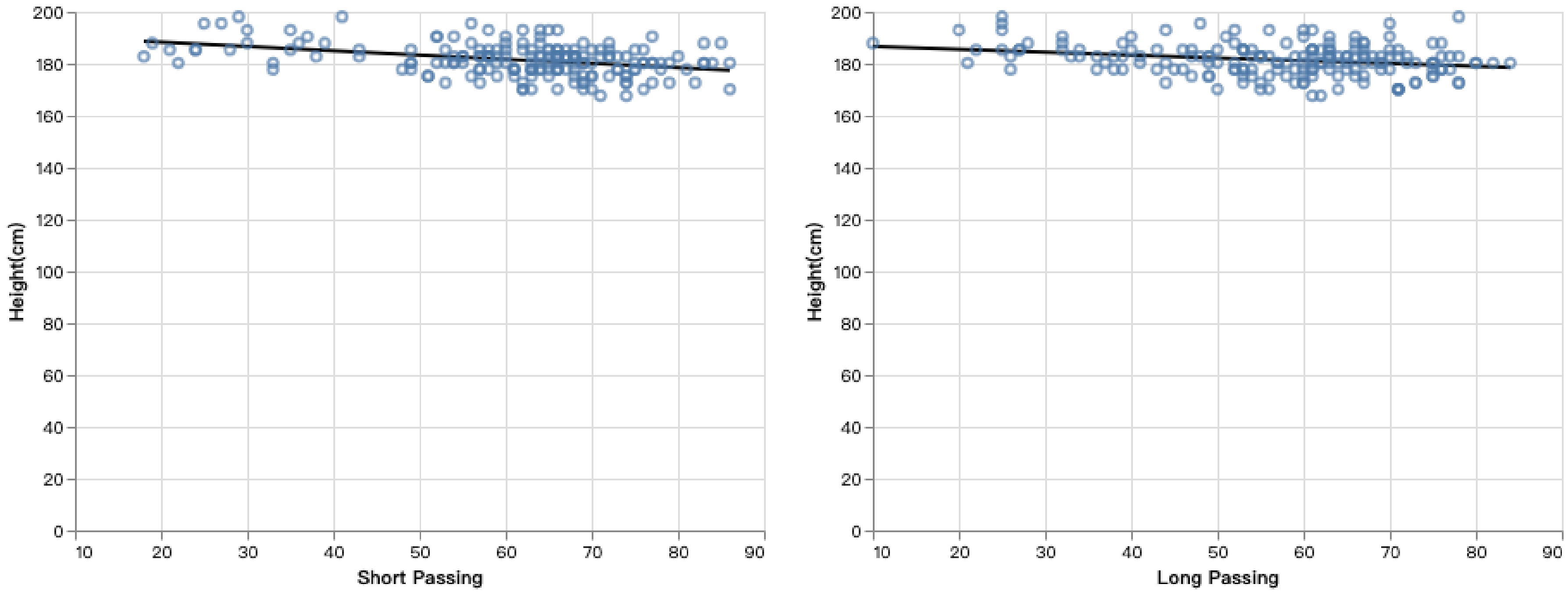
Graph #1 shows the league positions vs FIFA overall(the average overall of players on the team). Also, it demonstrates the relationship between real world league position and FIFA team overall attributes.



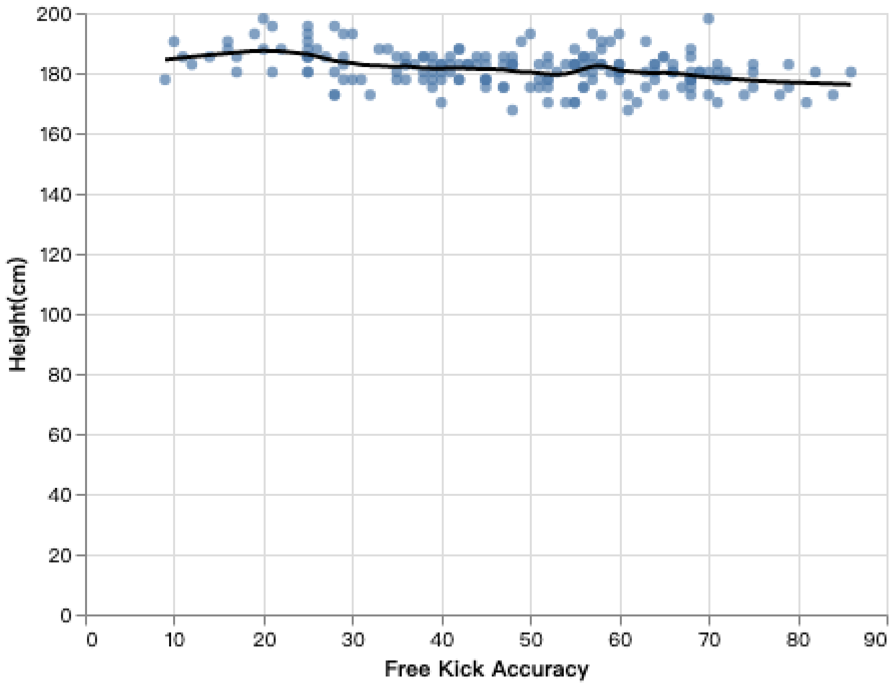
Graph #2 shows the relationship between team shooting and passing data in FIFA which related to real world league position.



Graph #3 is a data visualization that shows the average weight and height in the League. According to the graph, we calculate the average data and found that most of the player had a bmi around 23.5(Healthy range 18.5 - 24.9)



Graph #4 is deeper with the player attributes. The graph use regression method to show the average of height with high score and it demonstrates that player with 180cm had higher score with passing(short and long). In addition, the graph also predict that player with height between 182 - 180 had more efficient passing during the game.



Graph #5 use LOESS method to show the average of height with high score and it shows that player with 180cm had higher score with free kick accuracy.

---

3. Planned work

Here you should indicate your tentative ideas for your analysis. Don't worry, these aren't final -- you can always change your mind later or shift gears if they don't pan out. The objective is to have you start thinking ahead about what you'll do.

Questions

Please propose two focused questions that you plan to explore.

1. Are winning and scoring patterns the same across different European soccer leagues? (home advantage, goals scored, variability of top teams each year)
2. Are player and team attributes correlated with success? And if so, do the dominant attributes vary from league to league?

Proposed approaches

For each question, please describe an idea or two about how you might approach the question.

1. Collect data and visuals for each league and compare side by side.
2. Collect data and use regression model to predict the result then connect with real world data.

Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel* > *Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Generate PDF and submit to Gradescope