1. Remedies/Changes made any problems pointed out in assignment 2

There are two main issues with the initial dataset that we are working with to produce a database composed of multiple relations. The first issue is that some of the records in the dataset are not in English. This might cause an issue when a user is accessing the database as they may not be able to understand the language being used in the dataset but to resolve this issue we used a translate API to convert the text to English if possible.

The second issue at hand is that the dataset we are working with is composed of a single table and not separate relations. We processed the data into separate relations so that we can make logical relationships between the entities as described in our data dictionary and entity relationship diagram.

After having all separate relationships, we found that all of our entities are connected with the entity "Bio," so we tried to remove congestion. We found that the entity "Leagues" does not have enough attributes, so we combined the entities "League" and "Teams" into one: "Teams."

2. Database description (4-6 pg)

The database management system we are going to use is Sqlite.

For the database we will be using a Kaggle dataset on FIFA 18 players. FIFA is an association that is the international governing body of soccer (football). Kaggle is a website on which there are a multitude of datasets of different types. This dataset that we are using represents all the players in the FIFA 18 game that represents international teams, leagues for each country and many other smaller, independent teams. The main entity that we will be working with will be the players entity that represents a player and their statistics. Some of these statistics include attacking statistics, defense

statistics, position statistics, salary statistics, and other special characteristics dependent on the specific player at hand.

The Kaggle dataset that we are working with is a singular relation that contains players in FIFA 18 but we separated the dataset into multiple entities so that it could make more sense logically. As we know that each player has their own team, league, and countries, we created the table named 'Players_Bio' to connect every other tables. Also, we set player's FIFA ID as a primary key since it contains a unique value for each row of data and uniquely identify all table records also. The types of our data are text, numeric, date and time, hyperlinks for the pictures, and boolean. Most of the data is in numeric format and the data in table 'Trait' and 'Speciality' is entirely in boolean format.

Our entities are: player bios, overall statistics, attacking statistics, defense statistics, goalkeeper statistics, money statistics, teams, leagues, nationality, positions, traits, and specialties. It makes us organize similar types of numeric data such as sho(rating assigned to shooting (1-100)) and dri(rating assigned to dribbling (1-100)) for strikers, marking(rating assigned to marking (1-100)) and standing_tackle(rating assigned to standing tackle (1-100)) for defenders, and gk_diving(rating of how good a goal keeper is at diving (1-100)) and gk_handling(rating of how good a goalkeeper is at handling the ball (1-100)) for goalkeepers. Since the numeric data related to players' salary is a different type of data than score, we created the table called 'Money'. In the same manner, we created table 'Nationality' in order to separate players' nationality from different types of scores and salary data so that we can easily make a query related to the players' countries. Similar reasons applied to 'Positions', 'Traits', 'Speciality' and other tables. More detailed description of the list of tables and data is in data dictionary section.

Also, we created foreign key in each table except 'Players_Bio' table which is a references ID in players_bio table. Since foreign key is a field in the table that is primary key in another table and used to establish and enforce a link between the data in two tables, it makes sense for us to use ID in players_bio table for this project to connect each entities. These relations will be discussed further in detail in the data dictionary and the entity relationship diagram that shows how these entities relate to each other. The way that we generated each entities is simple but clear enough to not only the people who have professional knowledges in soccer but also to the normal people who are not into soccer very much by looking at the ER diagram below. By creating the relations, the users could attain the information of teams, leagues, and players more efficiently and effectively.

The database that will contain the separated entities and relations will allow a user of the database to see players, their statistics and their relations to other teams and leagues. Although there are many websites we could use to find about the players, but we found that none of them are really helpful, since they are mostly unorganized.

The main use for the database will be to see player statistics and what teams/leagues they are part of. Our target would be really wide. People who do not know soccer could use our database, to find the players' information, teams, leagues, etc. For example, the world-wide event, FIFA World Cup is coming, and we would like to come up with the database that could help people find the players they want to support. And, even soccer experts such as soccer agents or coaches can use the database to find out players they need, and we believe this database could help them scout players.

A user of this database can also query the relations to find players that match a certain criteria such as players with pass ratings greater than 90 or even more complicated queries. All the specific queries

users can run to find the specific players based on conditions would be explained in detail under section 6.

3. Data dictionary

Logical connections between them:

Player's have overall stats, attacking stats, defense stats, goal keeper stats, position stats, money stats, traits, and specialties.

Players play on teams, play in leagues, and play for countries.

Players_bio

text full_name: A player's full name

numeric ID: Primary Key: A player's FIFA ID

text club: Foreign key: The club team a player plays for.

text nationality: Foreign key: The country a player plays for.

DateTime birth date: Player's birthdate.

numeric age: Player's age.

numeric height cm: Player's height in cm.

numeric weight kg: Player's weight in kg.

hyperlink **photo**: Link to a photo of a player.

Overall_stat

numeric playerID: Foreign key: References ID in player_bio entity.

text **preferred_foot**: Player's preferred foot.

numeric overall: A player's overall rating.

numeric **potential**: A player's rating for potential.

numeric pac: Rating assigned to pace/speed.

text body type: Player's body type: normal, lean, stocky, etc

numeric weak foot: Rating of how good the player's weak foot is from 1 to 5 (the best)

text **international reputation**: Reputation of the players from 1 to 5 (the most famous)

numeric **stamina**: Rating of a player's stamina (1-100).

numeric **strength**: Rating of a player's strength (1-100).

numeric balance: Rating of a player's balance (1-100).

numeric **reactions**: Rating of a player's reactions (1-100).

numeric **heading accuracy**: Rating of a player's header accuracy (1-100).

numeric **interceptions**: Rating of how good a player's interceptions are (1-100).

numeric **positioning**: Rating of how good a player's positioning is (1-100).

numeric vision: Rating of how good a player's vision is (1-100).

numeric **penalties**: Rating of how good a player is at penalties (1-100).

numeric **composure**: Rating of how good a player's composure is (1-100).

Attacking_stat

numeric playerID: Foreign key: References ID in player bio entity.

numeric **sho**: Rating assigned to shooting (1-100).

numeric pas: Rating assigned to passing (1-100).

numeric **dri**: Rating assigned to dribbling (1-100).

numeric **crossing**: Rating assigned to crossing (1-100).

numeric **finishing**: Rating assigned to finishing (1-100).

numeric **short passing**: Rating assigned to short passing (1-100).

numeric volleys: Rating assigned to volleys (1-100).

numeric **dribbling**: Rating assigned to volleys (1-100).

numeric curve: Rating assigned to curve (1-100).

numeric free kick accuracy: Rating assigned to free kicking accuracy (1-100).

numeric **long passing**: Rating assigned to long passing (1-100).

numeric **ball control**: Rating assigned to ball control (1-100).

numeric acceleration: Rating assigned to acceleration (1-100).

numeric **sprint speed**: Rating assigned to sprint speed (1-100).

numeric agility: Rating assigned to agility (1-100).

numeric **shot power**: Rating assigned to shot power (1-100).

numeric **jumping**: Rating assigned to jumping (1-100).

numeric **shot power**: Rating assigned to acceleration (1-100).

numeric **long shots**: Rating assigned to long shots (1-100).

numeric **aggression**: Rating assigned to aggresion (1-100).

text work_rate_att: How active a player is on offense while not in possession of the ball (Low, Medium, High).

Defense stat

numeric playerID: Foreign key: References ID in player bio entity.

numeric def: Rating assigned to defense (1-100).

numeric **phy**: Rating assigned to physical (1-100).

text work_rate_def: How active a player is on defense while not in possession of the ball (Low, Medium, High).

numeric marking: Rating assigned to marking (1-100).

numeric standing_tackle: Rating assigned to standing tackle (1-100).

numeric sliding tackle: Rating assigned to sliding tackle (1-100).

Goalkeeper_stat

numeric playerID: Foreign key: References ID in player bio entity.

numeric **gk diving**: Rating of how good a goal keeper is at diving (1-100).

numeric **gk** handling: Rating of how good a goal keeper is at handling the ball (1-100).

numeric **gk** kicking: Rating of how good a goal keeper is at handling the ball (1-100).

numeric **gk positioning**: Rating of how good a goal keeper is at positioning (1-100).

numeric **gk reflexes**: Rating of how good a goal keeper's reflexes are (1-100).

Money

numeric playerID: Foreign key: References ID in player bio entity.

numeric eur wage: Player's wage in euros.

numeric eur value: Player's value in euros.

numeric eur_release_clause: Player's release buyback cost in euros.

Teams

numeric **playerID**: **Foreign key:** References ID in player_bio entity.

text club: A club team's name.

text league: Name of league.

hyperlink clublogo: Link to a photo of a club's team logo.

Nationality

numeric playerID: Foreign key: References ID in player_bio entity.

text country_name: Name of a country.

hyperlink **flag**: Link to a photo of a country's flag.

Positions

numeric playerID: Foreign key: References ID in player_bio entity.

numeric rs: Rating as a right sweeper.

numeric rw: Rating as a right winger.

numeric rf: Rating as a right forward.

numeric ram: Rating as a right attacking midfielder.

numeric **rdm**: Rating as a right defensive midefielder.

numeric rcb: Rating as a right center-back defender.

numeric rm: Rating as a right midfielder

numeric **rb**: Rating as a right full-back defender.

numeric rwb: Rating as a right wide-back defender.

numeric cf: Rating as a center forward.

numeric cam: Rating as a center attacking midfielder.

numeric cdm: Rating as a center defensive midfielder.

numeric cm: Rating as a center midfielder.

numeric **cb**: Rating as a center-back defender.

numeric ls: Rating as a left sweeper.

numeric lw: Rating as a left winger.

numeric If: Rating as a left forward.

numeric lam: Rating as a left attacking midfielder.

numeric **ldm**: Rating as a left defensive midefielder.

numeric **lcb**: Rating as a left center-back defender.

numeric Im: Rating as a left midfielder

numeric **lb**: Rating as a left full-back defender.

numeric lwb: Rating as a left wide-back defender.

numeric gk: Rating as a goal keeper.

Trait

numeric playerID: Foreign key: References ID in player bio entity.

boolean **chip shot trait**: Does the player have the chip shot trait?

boolean corner specialist trait: Does the player have the corner shot trait?

boolean diver trait: Does the player have the diver trait?

boolean finesse shot trait: Does the player have the chip shot trait?

boolean gk long throw trait: Does the player long throw as a goalkeeper?

boolean gk up for corners trait: Does the player up for corners as a goalkeeper?

boolean injury_free_trait: Is the player injury free?

boolean **injury prone trait**: Is the player prone to injury?

boolean leadership trait: Has the player been shown to be a leader?

boolean long_passer_trait: Does the player long pass?

boolean **long shot taker trait**: Does the player take long shots?

boolean one club player trait: Does the player only play for 1 club?

boolean playmaker_trait: Does the player make plays?

boolean **power_free_kick_trait**: Does the player power free kick?

boolean **power_header_trait**: Does the player power header?

Specialty

numeric playerID: Foreign key: References ID in player_bio entity.

boolean **speedster speciality**: Is strength this player's specialty?

boolean dribbler speciality: Is dribbling this player's specialty?

boolean engine_speciality: Is engine this player's specialty?

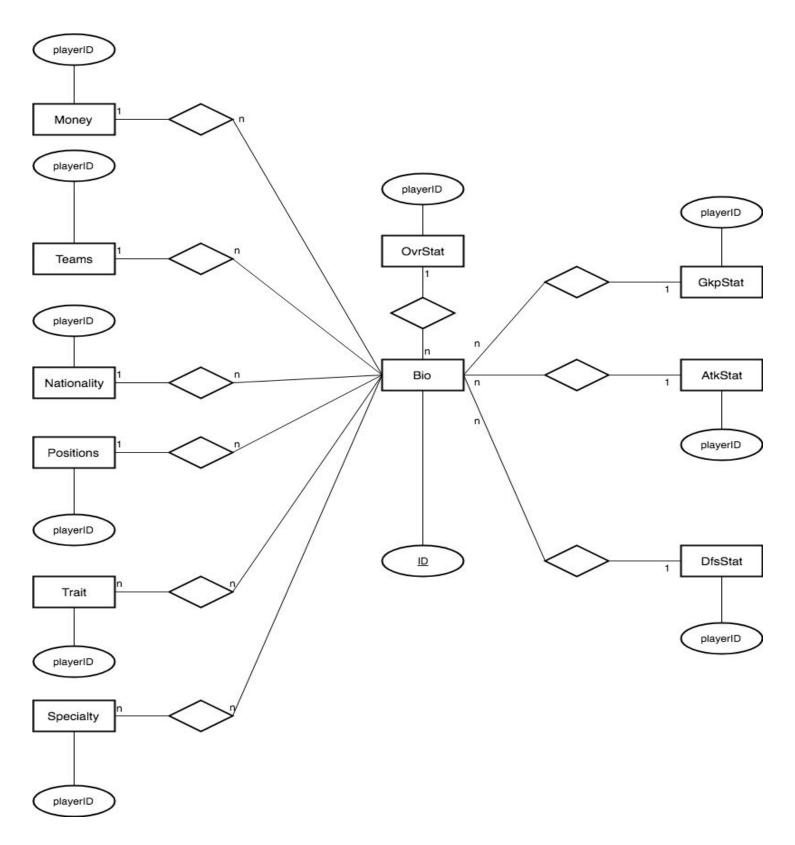
boolean distance_shooter_speciality: Is distance shooting this player's specialty?

boolean free kick specialist speciality: Is the player a free kick specialist?

boolean tackling speciality: Is tackling this player's specialty?

boolean strength_speciality: Is strength this player's specialty?

4. Relationships diagram



First of all, our database is fully normalized: every column has only one value, all of non-primary columns are dependent on the entire primary key, and there is non-key attribute that depends on another non-key attributes.

And, here is what each entity represent:

Players_bio: This entity contains the general information (that has nothing to do with soccer statistics) of the players.

| Field | Туре | Null | Key | Default | Extra |
|--|--|---|-----|---|-------|
| ID full_name club nationality birth_date age height_cm weight_kg photo | decimal(10,0) text text text text decimal(10,0) decimal(10,0) decimal(10,0) | NO YES YES YES YES YES YES YES | PRI | NULL NULL NULL NULL NULL NULL NULL NULL | |

Overall_stat: This entity contains the statistics any player could have regardless of their positions.

| Field | Туре | Null | Key | Default | Extra |
|--|--|------|-----|---|-------|
| playerID preferred_foot overall potential pac body_type weak_foot international_reputation stamina strength balance reactions heading_accuracy interceptions positioning vision penalties composure | decimal(10,0) text decimal(10,0) decimal(10,0) decimal(10,0) text decimal(10,0) decimal(10,0) | YES | MUL | NULL NULL NULL NULL NULL NULL NULL NULL | |

Attacking_stat: This entity contains the statistics how good the players are in attacking.

| + | Туре | Null | Key | Default | Extra |
|--|---|---|-----|---|-------|
| <pre>+</pre> | decimal(10,0) | YES | MUL | NULL NULL NULL NULL NULL NULL NULL NULL | |
| long_passing ball_control acceleration sprint_speed agility shot_power long_shots aggression work_rate_att | decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) text | YES | | NULL NULL NULL NULL NULL NULL NULL NULL | |

Defense_stat: This entity contains the statistics how good the players are in defending.

| + | Туре | + Null | Key | Default | Extra |
|---|--|---|------------------------|--|-------------|
| playerID def phy work_rate_def marking standing_tackle sliding_tackle | decimal(10,0) decimal(10,0) decimal(10,0) text decimal(10,0) decimal(10,0) decimal(10,0) | YES YES YES YES YES YES YES | MUL | NULL NULL NULL NULL NULL NULL NULL | |

Goalkeeper_stat: This entity contains the statistics mainly for the goalkeepers only.

| + Field | Туре | | | Default | Extra |
|---|---|-------------------------|-----|--|-------------|
| playerID gk_diving gk_handling gk_kicking gk_positioning gk_reflexes | decimal(10,0) decimal(10,0) decimal(10,0) text decimal(10,0) decimal(10,0) | YES YES YES YES YES YES | MUL | NULL NULL NULL NULL NULL NULL | |

Money: The entity contains the information that has to do with money.

| Field | Туре | Null | Key | Default | Extra |
|---|---|--------------------------|--------------|------------------------------|-------|
| playerID eur_wage eur_value eur_release_clause | decimal(10,0) decimal(10,0) decimal(10,0) text | YES YES YES YES | MUL | NULL NULL NULL NULL | |

Teams: This entity contains the information of leagues and clubs the players are belonged to.

| Field | | + | + | + | ++ |
|----------|---------------|------|-----|---------|-------|
| | Type | Null | Key | Default | Extra |
| | | + | + | | |
| playerID | decimal(10,0) | YES | MUL | NULL | |
| club | text | YES | | NULL | |
| league | text | YES | | NULL | |
| clublogo | text | YES | | NULL | |

Nationality: This entity contains the information what nationality the players have.

| Field | Туре | Null | Key | Default | Extra |
|--------------|---------------|------|-----|---------|-------|
| playerID | decimal(10,0) | YES | MUL | NULL | |
| country_name | text | YES | | NULL | |
| flag | text | YES | | NULL | |

Positions: The entity contains the information how effective the players are in different positions.

| Field | Туре | Null | Key | Default | Extra |
|----------|---------------|------|-----|---------|-------|
| playerID | decimal(10,0) | YES | MUL | NULL | |
| rs | decimal(10,0) | YES | | NULL | |
| rw | decimal(10,0) | YES | | NULL | |
| rf | decimal(10,0) | YES | | NULL | |
| ram | decimal(10,0) | YES | | NULL | |
| rdm | decimal(10,0) | YES | | NULL | |
| rcb | decimal(10,0) | YES | | NULL | |
| rm | decimal(10,0) | YES | , | NULL | |
| rb | decimal(10,0) | YES | | NULL | |
| rwb | decimal(10,0) | YES | | NULL | |
| cf | decimal(10,0) | YES | | NULL | |
| cam | decimal(10,0) | YES | | NULL | |
| cdm | decimal(10,0) | YES | | NULL | |
| cm | decimal(10,0) | YES | | NULL | |
| cb | decimal(10,0) | YES | | NULL | |
| ls | decimal(10,0) | YES | | NULL | |
| lw | decimal(10,0) | YES | | NULL | |
| lf | decimal(10,0) | YES | | NULL | |
| lam | decimal(10,0) | YES | | NULL | |
| ldm | decimal(10,0) | YES | | NULL | |
| lcb | decimal(10,0) | YES | | NULL | |
| lm | decimal(10,0) | YES | | NULL | |
| lb | decimal(10,0) | YES | | NULL | |
| lwb | decimal(10,0) | YES | | NULL | |
| gk | decimal(10,0) | YES | | NULL | |

Trait: This entity contains the information of unique traits the players have, and these traits are especially useful for scouts.

| Field | Туре | Null | Key | Default | Extra |
|---|---|---|-----|---|-------------|
| playerID chip_shot_trait corner_specialist_trait diver_trait finesse_shot_trait gk_long_throw_trait gk_up_for_corners_trait injury_free_trait injury_prone_trait leadership_trait long_passer_trait long_shot_taker_trait one_club_player_trait playmaker_trait power_free_kick_trait power_header_trait | decimal(10,0) | YES | MUL | NULL NULL NULL NULL NULL NULL NULL NULL | |

Specialty: The entity contains the information where the players are specialized in. (i.e. who is taking free-kicks)

| + Field | + Type | Null | Key | Default | Extra |
|--|---|---|-----|---|-------------|
| playerID speedster_speciality dribbler_speciality engine_speciality distance_shooter_speciality free_kick_specialist_speciality tackling_speciality strength_speciality | decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) decimal(10,0) | YES | MUL | NULL NULL NULL NULL NULL NULL NULL NULL | |

And, here is why they are linked this way:

The playerID column is the only unique id that can differentiate each player, so playerID is the Key attribute for each entity. And, playerID is one of general information, and thus, it is appropriate to belong to the entity: "Players_bio." So, we made PlayerID as a primary key for "Players_bio," and let this be a foreign key for all other entities.

5. Sample queries and results

These three queries are for people who just get started to learn soccer.

Query 1: I am interested in finding ten players who are in specific teams and have the same nationality(for example, Spain).

SELECT Bio.full_name, Bio.nationality, Bio.club, Bio.photo FROM Bio
WHERE Bio.nationality = 'Spain' LIMIT 10;

David De Gea Quintana|Spain|Manchester United|https://cdn.sofifa.org/18/players/193080.png
Sergio Ramos García|Spain|Real Madrid CF|https://cdn.sofifa.org/18/players/155862.png
Thiago Alcântara|Spain|FC Bayern Munich|https://cdn.sofifa.org/18/players/189509.png
David Josué Jiménez Silva|Spain|Manchester City|https://cdn.sofifa.org/18/players/168542.png
Gerard Piqué Bernabeu|Spain|FC Barcelona|https://cdn.sofifa.org/18/players/152729.png
Andrés Iniesta Luján|Spain|FC Barcelona|https://cdn.sofifa.org/18/players/41.png
Francisco Román Alarcón Suárez|Spain|Real Madrid CF|https://cdn.sofifa.org/18/players/197781.png
Sergio Busquets Burgos|Spain|FC Barcelona|https://cdn.sofifa.org/18/players/189511.png
Diego da Silva Costa|Spain|Chelsea|https://cdn.sofifa.org/18/players/179844.png
Javier Martínez Aginaga|Spain|FC Bayern Munich|https://cdn.sofifa.org/18/players/177610.png

Query 2: Which country has the most players at foreign clubs?

SELECT Bio.nationality, count(DISTINCT Teams.club) as club_count FROM Bio, Teams

WHERE Bio.ID = Teams.playerID

GROUP BY Bio.nationality

ORDER BY club_count DESC LIMIT 1;

Brazil|221

Query 3: Who are the highest-paid player for each team (top 10)

SELECT Bio.full_name, Bio.club, Money.eur_wage
FROM Bio, Money
WHERE Bio.ID = Money.playerID AND (Bio.club, Money.eur_wage) IN
(SELECT Bio.club, max(Money.eur_wage)
FROM Bio, Money
WHERE Bio.ID = Money.playerID
GROUP BY Bio.club)
ORDER BY Money.eur_wage DESC LIMIT 10;

C. Ronaldo dos Santos Aveiro|Real Madrid CF|565000

Lionel Messi|FC Barcelona|565000

Robert Lewandowski|FC Bayern Munich|355000

Sergio Agüero|Manchester City|325000

Eden Hazard|Chelsea|295000

Neymar da Silva Santos Jr.|Paris Saint-Germain|280000

Gonzalo Higuaín|Juventus|275000

Alexis Sánchez|Arsenal|265000

Mesut Özil|Arsenal|265000

Zlatan Ibrahimović|Manchester United|240000

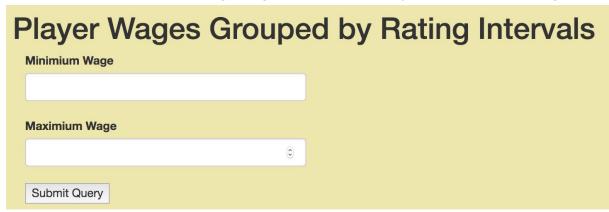
This query is for agents, scouts, and coaches.

Query 4: We want a defender, aged between 23 - 28 year-old who is likely to get injured

SELECT B.full_name, D.def, B.age
FROM Defense_stat D, Bio B, Trait T
WHERE D.playerID = B.ID AND B.ID = T.playerID
AND B.age >= 23 and B.age <= 28 AND T.injury_prone_trait != 0
ORDER BY D.def DESC limit 1;

6. Sample forms

Form 1: For a person interested in drafting the best players for his dream team but has a budget, a user can input a min and max wage range to find the best player based on their rating.

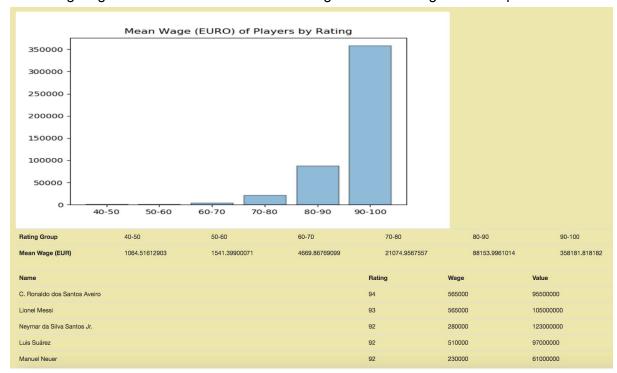


Form 2: For Query 1, we include simple select drop down form to allow a person to select the nationality of the players they are interested in.



7. Sample reports (at least 2 reports)

Report 1: For our first report, a user is able to set a wage-range for the players they want to see, with all players being displayed by default. A graph displaying the mean range based on ratings (a bin size of 10), along with a table that also displays the group means. Finally, a table of players, sorted in descending rating, then by ascending wage to allow a someone like a manager to find the highest rated person for the cheapest wage.



Report 2: Users are able to select their country from the select form. Summary statistics, along with a table with all players from the selected country are generated.

