Creating the 2018 NCAA Men’s basketball database:

**Summary**

The first step in designing the database was to determine the purpose of the database. Then determine the tables and columns needed to fulfill the purpose of the database. Lastly, determine how each table and column should relate to the other in order to fulfill the database design requirements.

**Database Purpose**: The purpose of the 2018 NCAA Men’s basketball database is to store player statistics and team statistics

**Determine Tables and Columns** : In order to determine the tables and columns the database designer should refer to the requirements which state the data to be stored in the database:

**Data to be stored in database:**

Player person data -

* Player name
* Player’s team
* Players position
* Player height
* Player weight
* Date of Birth

Player game data -

* Minutes played per game
* Field goals made per game
* Field goals attempted per game
* Three -point shots attempted per game
* Three-point shots made per game
* Free throws attempted per game
* Free throws made per game
* Offensive rebounds per game
* Defensive rebounds per game
* Assists per game
* Steals per game

Team data -

* Team name
* Team region
* Team ranking

Team game data -

* Minutes played per game
* Field goals made per game
* Field goals attempted per game
* Three -point shots attempted per game
* Three-point shots made per game
* Free throws attempted per game
* Free throws made per game
* Offensive rebounds per game
* Defensive rebounds per game
* Assists per game
* Steals per game

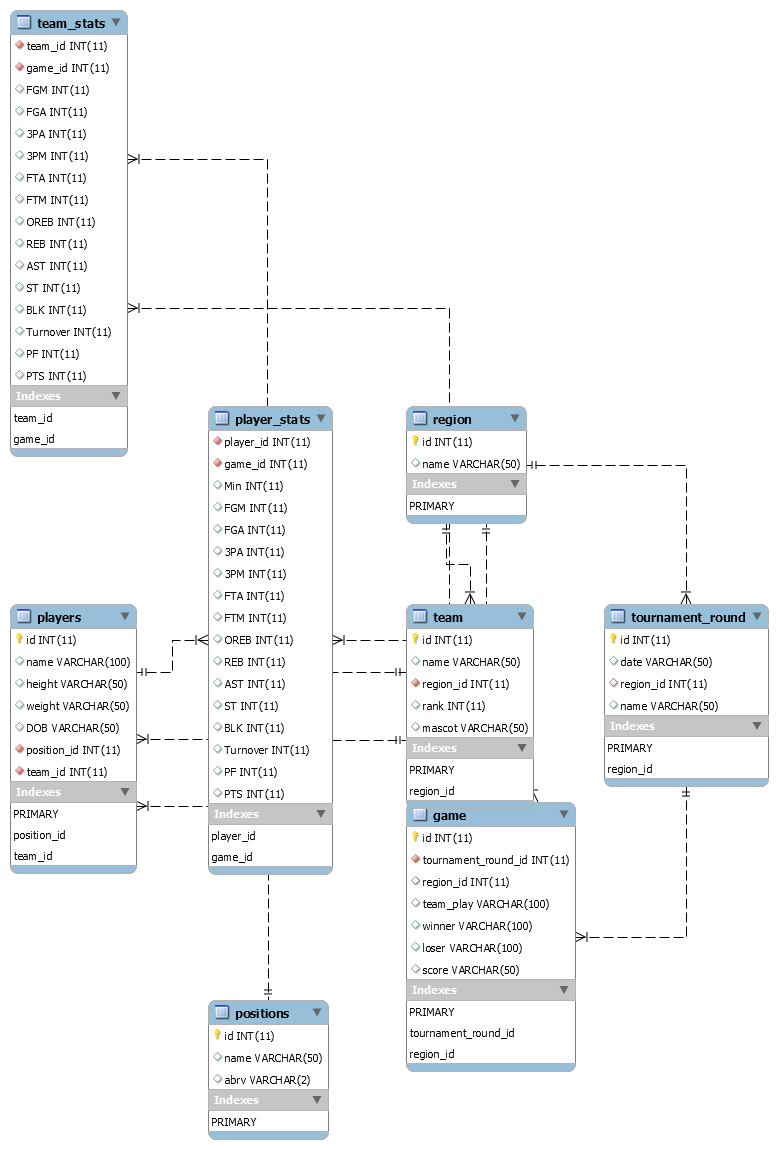
**Determine relationship of database**: Refer to requirements to determine key insights and trends of interest to client.

**Insights of Interest to Client**:

* Shooting percentage per game
* Shooting percentage per player
* Turnover percentage per game
* Turnover percentage per player
* Efficiency percentage: Efficiency = (points/possessions) \* 100

**Final output:** mySQL workbench was used to create a database from scratch using sql scripts. Full code can be found in ‘database/ ncaam2018\_elite\_8 / create\_ncaam2018\_db.sql’

**ER Diagram –**



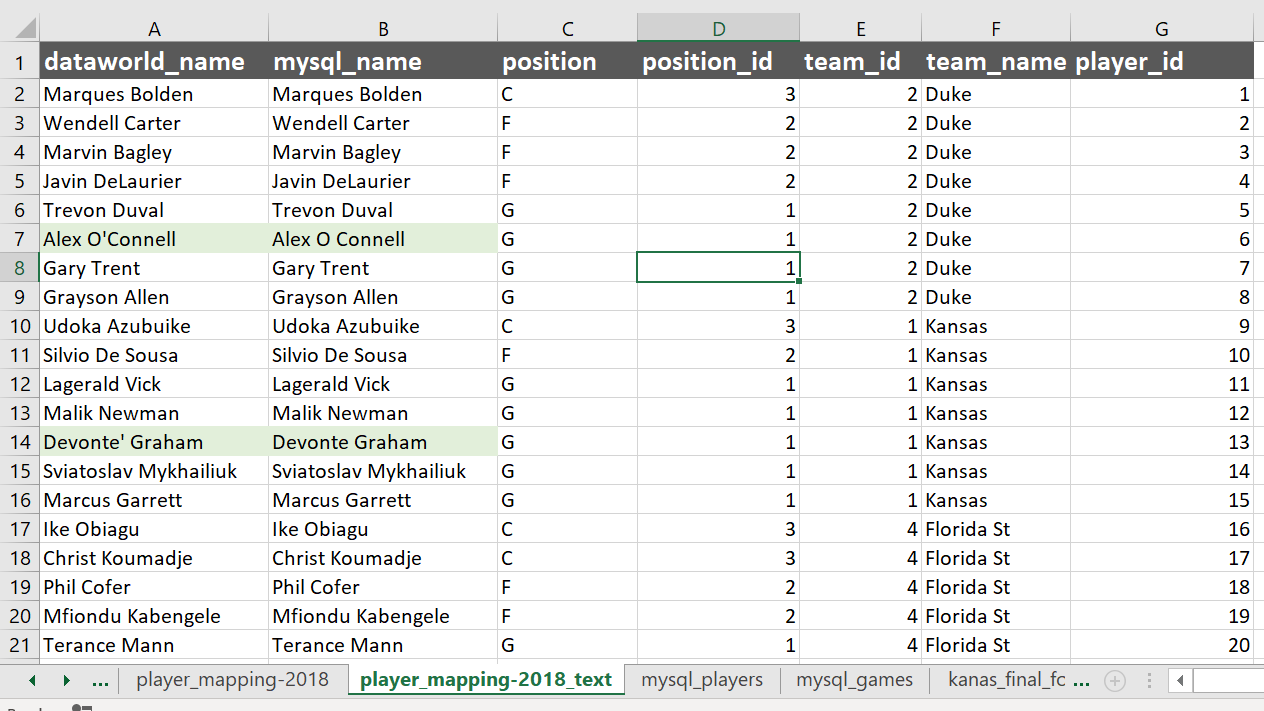
Loading NCAA Men’s basketball database:

**Summary**:

The greatest challenge to loading the database was transforming external data sources to meet the structure requirements of the database. The external data sources were scraped from the NCAA website and the raw csvs did not have the full columns or information needed to load the csvs to the database. The raw csvs did not have the team name, winner or loser information available. Therefore, a python script was created to transform the raw data using jupyter notebook. See data\_transform\_load.iynb for full code.

**Finding Resolutions**: A common challenge of loading data from external sources is a different naming convention used to refer to the same data point, or the limitations of special characters. For example, Duke University player Alex O'Connell. mySQL had troubles loading the players name. Additionally, mySQL used the name ‘Mount St. Mary’ while external data sources used names like ‘Mount Saint Mary’. To overcome these challenges, a mapping document was created. The mapping document outlines and tracks the names used by each data source and how item is mapped to internal database.

**Snapshot of mapping document**:



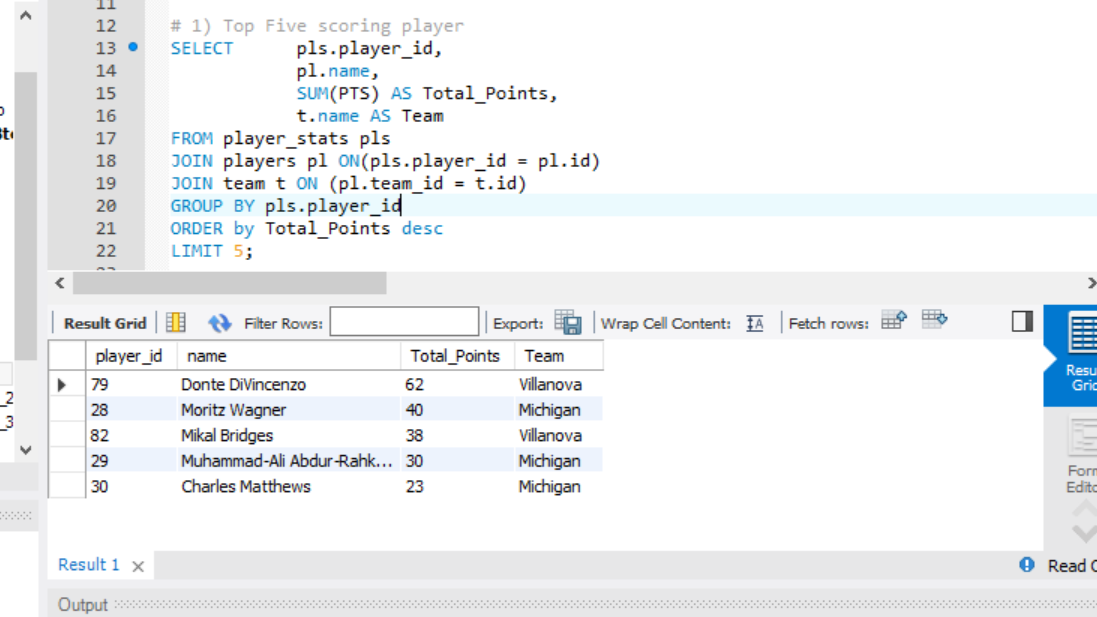
**Providing Insight**:

Once the hard work of designing and loading the database is complete, sql queries can be used to probe the database for helpful insight. For example, determine each teams most, least efficient game and the opponent; determine the number of times a lower ranking team won and which categories did the lower ranking team outperform a higher ranking team; who was the conference leading three-point shooter, shot blocker, and free throw shooter. The possibilities are endless. Sample queries were provided to the client in the ‘sql\_queries’ folder as a good starting point.

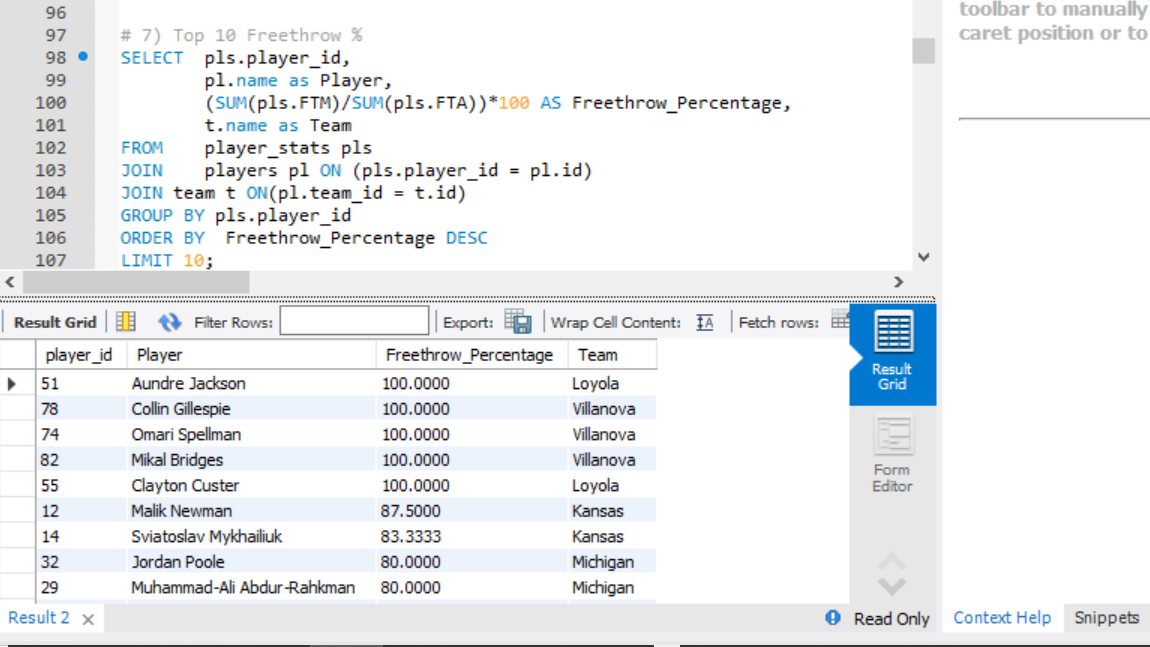
Mining the Database for Insight:

With such valuable data now stored in the client’s database, the next step is mining the database for hidden gems. The file ‘performance\_queries.sql’ is a full list of queries provided to breakdown team performance and individual player performance. Below is a screenshot two of the top five scoring players in the final four and championship games and the top ten free-throw percentage. For a full list of queries available see ‘performance\_queries\_. xlsx

Top 5 scoring players



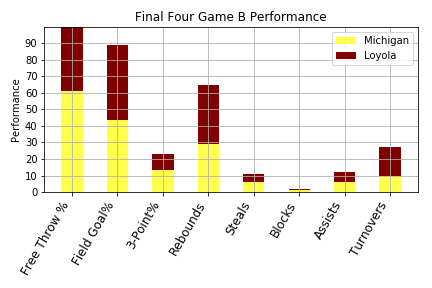
Top 10 Free-throw percentage

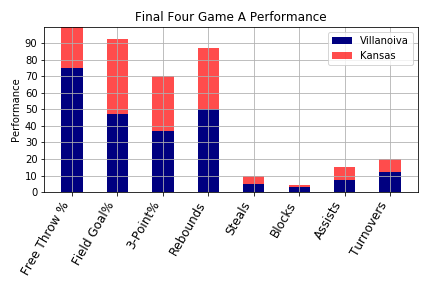


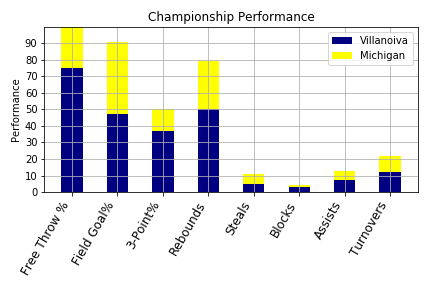
to assist with digesting team performance. orward would great graphical visualizations of the data. For example, creating a bubble chart showcasing different shot selections to determine if there is a concentration of shots taken for specific categories. Additionally, attempts and makes can be use different colors or symbols. This can help guidance to coaches. A line chart which follows the number of turnovers for each half or quarter can help coaches visualize if there is a turnover trend. Whether it is visualizing how many times a team faced a team and the game outcomes, or identifying areas of strength and areas for improvement, the opportunities are possible due to the well-designed database.

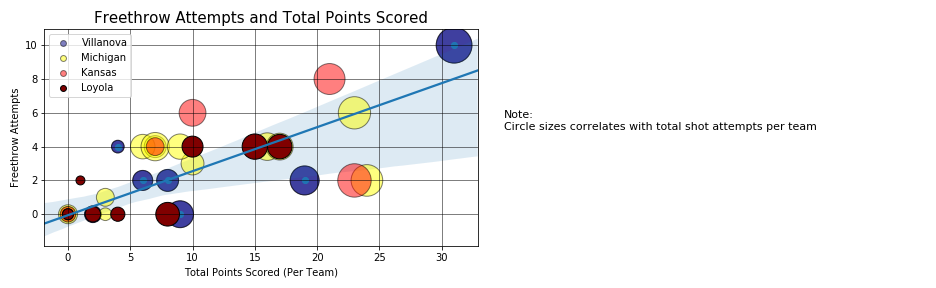
Visualization:

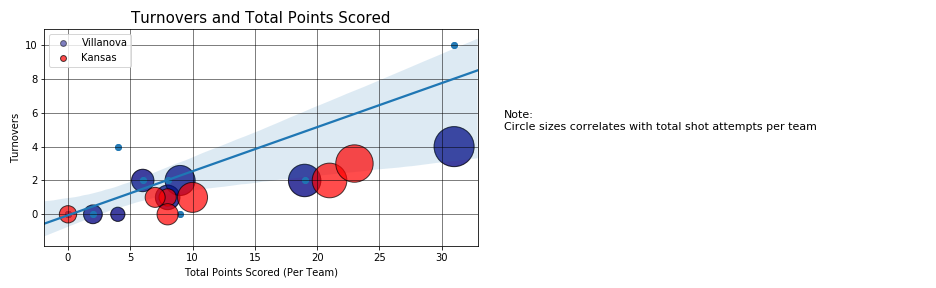
The final step is a visual analysis of a graphical visualizations of the data. Final Four game performance, Championship game performance, the relationship between of free-throw attempts and overall scoring, the relationship between turnovers and overall scoring, minutes played and scoring productivity is provided in the jupyter notebook ‘stats\_visualzation’. Graphs are provided below:

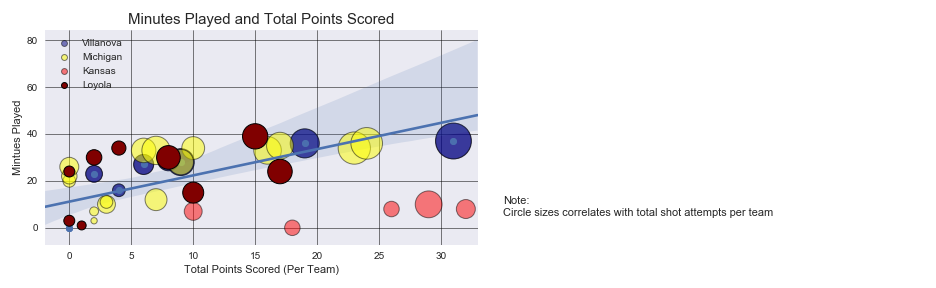












**A look forward**:

With more time and resources available, an area of improvement would be conducting tests for data validation and reasonability. For example, pulling the statistics and validating that each player’s aggregated statistics matches the team’s statistics and matches the overall score.