The purpose of this document is to explain any decisions that were made in the coding process which may not have been clearly outlined in the code itself. It is not necessary to read through this entire document, but if you are curious as to why I did something then the explanation can be found here.

**Importing the Dataset and Filtering out Irrelevant Features**

* Data was downloaded from evolving-hockey.com
* **Creating a “Player\_Season” Column:** The creation of the “Player\_season” column, which is simply the players name and then the season that they played in put together, is for the purpose of having a unique code for each player season. Many players have played for multiple seasons within this dataset, so there was no unique code for many of the specific columns in the dataset.
* **What data was downloaded?**
  + **Forwards only:** It is reasonable to test the idea of chemistry with only forwards, only defenseman, or both. All would give different perspectives on how players fit together, but for the scope of this analysis I wanted to start by looking only at forwards.
  + **5v5 only:** It would be interesting to examine chemistry on the power play or other game states, but they are vastly different play styles so they should be examined in isolation. 5v5 is the most common and important game state, so that is what was looked at in this analysis.
  + **Per 60 minutes rates, not totals**
  + **Last 10 seasons (2011-12 season to the 2020-21 season):** This gives us a strong sample size of data without going too far back to where you need to start considering hockey's stylistic changes to a massive degree.
  + **Minimum TOI of 500 minutes:** While using rate stats, it’s important to get rid of low sample size players who may have extreme results. 500 minutes is arbitrary, but it is approximately how much any semi-regular player in a lineup would play over the course of a season.
    - Note: The 2012-13, 2019-20 and 2020-21 were all shortened seasons, but 500 minutes was still extremely achievable in all of these years and we still want to maintain a strong sample size for each player.
* After applying these filters, we are left with 3301 observations (player seasons).
* The best data would have been micro stats (e.g. zone entry data, shot assists, etc.) because that describes a players style better than macro stats. However, the NHL only tracks some basic individual stats, so for the purpose of this analysis we will work with what we have, keeping in mind that the player clustering is far from perfect.
* **Features Used in the Model:**
  + **G/60:** Goal scoring ability.
  + **A1/60, A2/60:** Playmaking ability.
  + **iCF/60, ixG/60:** Volume and quality shot generating ability.
  + **Fsh%\_Over\_Exp:** Shooting talent.
  + **iBLK/60, TAKE/60:** These are the only defensive micro stats that the league tracks.
  + **GIVE/60:** Turnover rate.
  + **iHF/60:** Physicality.
  + **iHA/60:** Could be an indicator of a lot of different things, including willingness to go to “the dirty areas”, a lack of escapability, or a number of other things.
  + **FO±/60:** Faceoff ability.
  + **xEVD\_GAR/60:** I tried to avoid using macro stats because they describe how good a player’s on ice results are, not how they achieved those results. But I made an exception here so I could incorporate more defensive data.
* **Features Filtered Out from the Data and Omitted from the Model:**
  + **Player, Team, and Season:** These are obviously not relevant to a players style, but will be added in later to see which players are in which cluster.
  + **Position:** Originally I did include position by using dummy variables, but I found that the clustering model severely overweighed a players position, essentially making each cluster just a different position. I do believe that a players position is relevant to their players chemistry together, but that should be examined separately in a future study.
  + **GP, TOI:** The total ice time a player plays is not relevant to their player type, remember we don’t care about how good the player is (TOI may be a proxy for that), we only care about the type of player so their TOI doesn’t matter.
  + **Points/60:** I am using goals, primary assists, and secondary assists as features so using total points would just be double counting.
  + **iSF/60, iFF/60:** I am using corsi instead.
  + **Sh%, FSh%:** I am using shooting percentage over expected instead.
  + **xFSH%:** I am using iXG/60 instead.
  + **All penalty data:** While the propensity of a player to draw or take penalties may impact the way that a line plays at 5v5, I don’t think that is likely or strong enough to include as a feature, it would likely just add noise.

**PreProcessing the Data**

* **Scaling the data:** We are scaling the data because, as I understand it, machine learning uses the variance of data points as one way to weigh them. If data points are presented on different scales, they will have impacts on the model outside of their actual importance. So putting all the data on the same scale should help weigh features properly.

**Building the Clustering Model**

* **Using the “Elbow Method” to Determine the Amount of Clusters:** Using the “elbow method” is a way to determine the appropriate amount of clusters that we need to use. For this specific analysis, less than 3 clusters or more than 8 clusters does not make much sense, as there are obviously more than 2 player types and probably less than 8 clearly defined player types.
* **Choosing 5 clusters:** The “elbow method” essentially showed that any number between 3 and 8 is appropriate for this analysis. Choosing 5 clusters was somewhat subjective, but it seemed like an appropriate amount given the amount of unique player types in the NHL that still gives us a big enough sample size for each cluster.

**Testing Cluster Chemistry - Player Pairs**

* Now that we have each player in a cluster, we can look at how well each cluster played with each other.
* **What data is being used to test this?**
  + This data comes from the “teammate tool” on Evolving-Hockey.com.
  + **Both players are forwards:** We are obviously only looking at forwards as this is the scope of this analysis.
  + **Last 10 seasons**
  + **Regular Season:** This gives us a much bigger sample size than looking at the playoffs.
  + **60 Minutes as the Minimum TOI together** This is the default setting.
* After the filters, we are left with 27,671 unique combinations of players playing together.
* **Why do we only care about “Player” and not “Teammate”?** Because each player pair is counted twice, once with each as the “Player” and once as the “Teammate”, so if we calculate both then we end up double counting.