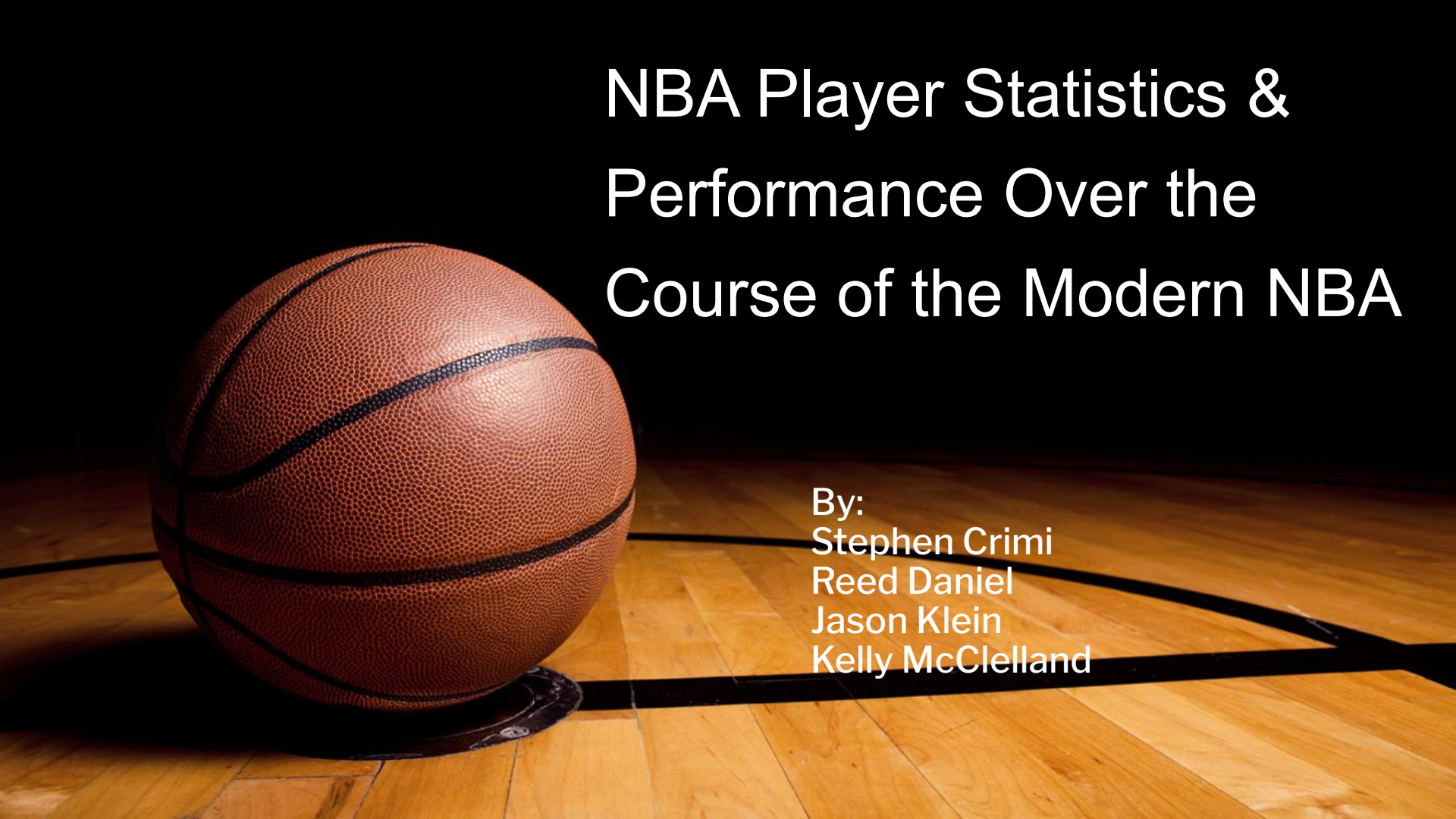


NBA Player Statistics & Performance Over the Course of the Modern NBA

By:
Stephen Crimi
Reed Daniel
Jason Klein
Kelly McClelland

Topic:

The relationship between NBA player statistics and performance over the course of the modern NBA.

Reason for selecting the topic:

- To explore statistical data to determine if the NBA has a baseline for its athletes
- How that baseline weighs against the top performers throughout the decades
- Predict the number of games a player would play per season.

Source Data

- Our source data is from Kaggle and includes NBA Players Stats since 1950 in the form of .csv files.
- For purpose of our analysis we will only focus on the years 1980 to the present as that is when the “modern” NBA began.

File Name	Number of Rows	Number of Columns
Player.csv	3,922	8
Season_Stats.csv	24,690	52
Player_Data.csv	4,550	8
NBA_Players_AllStars_All.csv	943	7

File Name	Number of Rows	Number of Columns
player_df.shape	3,919	7
seasons_df_shape	18,297	52
per_game_df.shape	18,297	11

Questions to Answer

- Does the NBA look the same decade by decade in terms of performance?
- What does a prototypical player look like in each decade?
  - How has that changed over time?
- Can you predict whether an NBA player can be an All-Star?
  - What are the most important stats in terms of determining an NBA All-Star?

Project and Communication Protocols

- Team Communication
  - Slack
  - Teams
  - Email
- GitHub
  - The main branch:
    - Code
      - Perform exploratory analysis & machine learning
    - README.md includes:
      - Description of the communication protocols
      - Outline of the project (this may include images, but should be easy to follow and digit).
    - Individual Branches
    - Topic Branches

Project and Communication Protocols (con’td)

Responsibility Legend	Symbol	Team Member	Systems	Phase
The team member in the square role will be responsible for the repository	□	K. McClelland	GitHub Teams Calendar	I - IV
The member in the triangle role will create a mockup of a machine learning model. This can even be a diagram that explains how it will work concurrently with the rest of the project steps.	Δ	S. Crimi	Machine Learning	I - IV
The member in the circle role will create a mockup of a database with a set of sample data, or even fabricated data. This will ensure the database will work seamlessly with the rest of the project	○	J. Klein	Database	I - IV
The member(s) in the X Roll will decide which technologies will be used for each step of the project	X	S. Crimi J. Klein R. Daniel	Machine Learning Database Dashboard	I - IV I - IV II - IV

Technologies Used

Data Clean & Analysis

Pandas

Matplotlib

Numpy

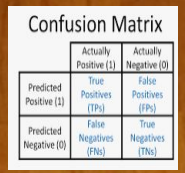
Database Storage

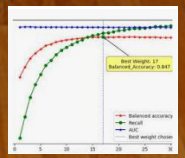
Dashboard

Lambda

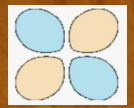
FSD & ERD

Machine Learning





Balance Accuracy Score



Classification Report Imbalanced

Results & visualizations displayed

Google Slides

READMEs

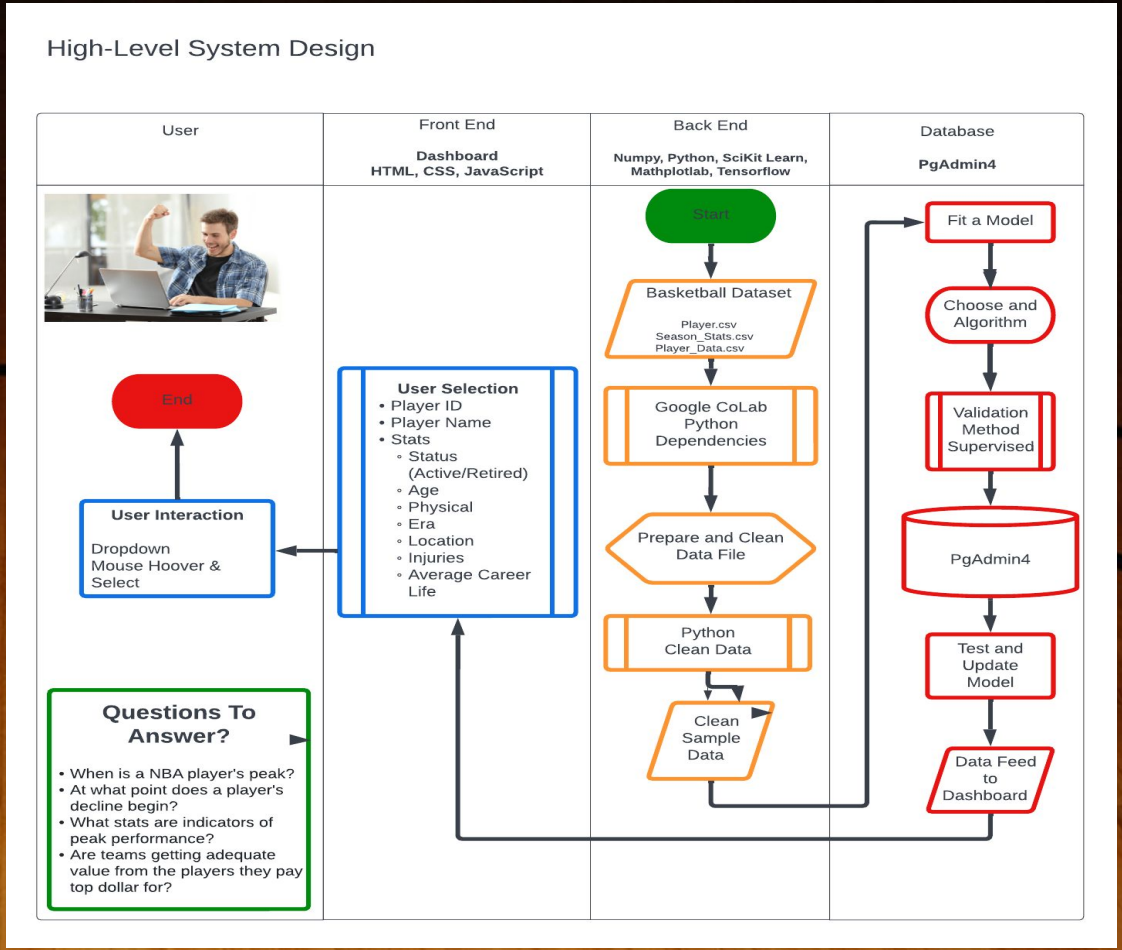
Tools

Systems	Tools
Data Cleaning & Analysis	Pandas, Numpy, MathPlotLib
ERD & FSD	Power BI, Lucid Charts, QuickDBD
Database	PgAdmin4, AWS
Machine Learning	SciKitLearn, Jupyter Notebook, Supervised Learning Model
Dashboard	HTML, CSS, JavaScript, Tableau, GitHub Pages

ERD & FSD

FSD

High-Level FSD



Database

Database (cont'd)

- Perform data cleaning and output ETL documents
- Load created schema into PostgresSQL via pgAdmin
- Upload data into PostgreSQL database
- Create AWS RDS and connect to PostgreSQL
- Connect AWS RDS to Python with SQLAlchemy

Clean data and end of ETL

```
# merge_players DF to csv file
merge_players.to_csv('players_clean.csv')
#files.download('players_clean.csv')
# modern_season_stats DF to csv file
merge_modern_season_stats.to_csv('modern_season_stats_clean.csv')
files.download('modern_season_stats_clean.csv')

# per_game_stats DF to csv file
per_game_stats.to_csv('per_game_stats_clean.csv')
files.download('per_game_stats_clean.csv')
```

```
merge_players.sample(2)
```

	Player	height	weight	born	year_start	year_end	position
3187	Brandon Roy	78	214	1984	2007.0	2013.0	G
1447	Billy Reid	77	190	1957	1981.0	1981.0	G

```
merge_modern_season_stats.head(2)
```

	Year	Player	Pos	Age	Tm	G	GS	MP	PER	TS%	...	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	id	AllStar
0	1980	Kareem Abdul-Jabbar	C	32	LAL	82	0	3143	25.3	0.639	...	696	886	371	81	280	297	216.0	2034	0	1
1	1980	Tom Abernethy	PF	25	GSW	67	0	1222	11.0	0.511	...	129	191	87	35	12	39	118.0	362	1	0

2 rows × 52 columns

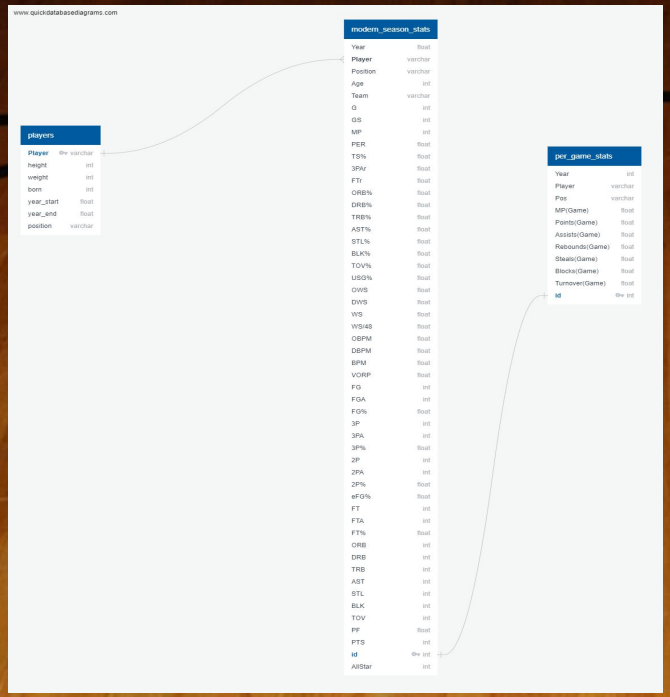
```
per_game_stats.head(2)
```

	Year	Player	Pos	MP(Game)	Points(Game)	Assists(Game)	Rebounds(Game)	Steals(Game)	Blocks(Game)	Turnover(Game)	id
5727	1980	Kareem Abdul-Jabbar	C	38.3	24.8	4.5	10.8	1.0	3.4	3.6	0
5728	1980	Tom Abernethy	PF	18.2	5.4	1.3	2.9	0.5	0.2	0.6	1

Database (cont'd)

Database ERD

- Load created schema into Postgres via pgAdmin



schema ERD

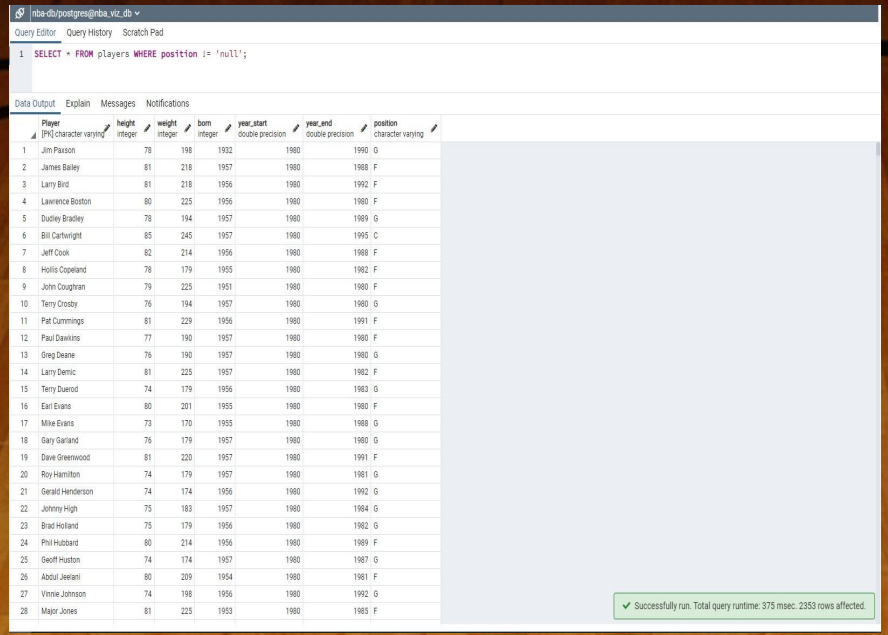
Sample of schema uploaded into Postgres

```sql
-- Exported from QuickDBD: https://www.quic
-- Link to schema: https://app.quickdatabas
-- NOTE! If you have used non-SQL datatypes


-- Players table
CREATE TABLE "players" (
    "Player" varchar   NOT NULL,
    "height" int   NOT NULL,
    "weight" int   NOT NULL,
    "born" int   NOT NULL,
    "year_start" float   NOT NULL,
    "year_end" float   NOT NULL,
    "position" varchar   NOT NULL,
    CONSTRAINT "pk_players" PRIMARY KEY (
        "Player"
    )
);

-- modern_season_stats table
CREATE TABLE "modern_season_stats" (
    "Year" float   NOT NULL,
    "Player" varchar   NOT NULL,
    "Position" varchar   NOT NULL,
    "Age" int   NOT NULL,
    "Team" varchar   NOT NULL,
    "G" int   NOT NULL,
    "GS" int   NOT NULL,
    "MP" int   NOT NULL,
    "PER" float   NOT NULL,
    "TS%" float   NOT NULL,
    "3PAr" float   NOT NULL,
    "FTr" float   NOT NULL,
    "ORB%" float   NOT NULL,
    "DRB%" float   NOT NULL,
    "TRB%" float   NOT NULL,
    "AST%" float   NOT NULL,
    "STL%" float   NOT NULL,
    "BLK%" float   NOT NULL,
    "TOV%" float   NOT NULL,
    "USG%" float   NOT NULL,
    "OWS" float   NOT NULL,
    "DWS" float   NOT NULL,
    "WS" float   NOT NULL,
    "WS/48" float   NOT NULL,
    "OBPM" float   NOT NULL,
    "DBPM" float   NOT NULL,
    "BPM" float   NOT NULL,
    "VORP" float   NOT NULL,
    "FG" int   NOT NULL,
    "FGA" int   NOT NULL,
    "FG%" float   NOT NULL,
    "3P" int   NOT NULL,
    "3PA" int   NOT NULL,
    "3P%" float   NOT NULL,
    "2P" int   NOT NULL,
    "2PA" int   NOT NULL,
    "2P%" float   NOT NULL,
    "eFG%" float   NOT NULL,
    "FT" int   NOT NULL,
    "FTA" int   NOT NULL,
    "FT%" float   NOT NULL,
    "ORB" int   NOT NULL,
    "DRB" int   NOT NULL,
    "TRB" int   NOT NULL,
    "AST" int   NOT NULL,
    "STL" int   NOT NULL,
    "BLK" int   NOT NULL,
    "TOV" int   NOT NULL,
    "PF" float   NOT NULL,
    "PTS" int   NOT NULL,
    "id" int   NOT NULL,
    "AllStar" int   NOT NULL,
    CONSTRAINT "pk_modern_season_stats" PRIMARY KEY (
        "id"
    )
);

-- per_game_stats table
CREATE TABLE "per_game_stats" (
    "Year" int   NOT NULL,
    "Player" varchar   NOT NULL,
    "Pos" varchar   NOT NULL,
    "MP(Game)" float   NOT NULL,
    "Points(Game)" float   NOT NULL,
    "Assists(Game)" float   NOT NULL,
    "Rebounds(Game)" float   NOT NULL,
    "Steals(Game)" float   NOT NULL,
    "Blocks(Game)" float   NOT NULL,
    "Turnover(Game)" float   NOT NULL,
    "id" int   NOT NULL,
    CONSTRAINT "pk_per_game_stats" PRIMARY KEY (
        "id"
    )
);

ALTER TABLE "modern_season_stats" ADD CONSTRAINT "fk_modern_season_stats_Player" FOREIGN KEY("Player")
REFERENCES "players" ("Player");

ALTER TABLE "modern_season_stats" ADD CONSTRAINT "fk_modern_season_stats_id" FOREIGN KEY("id")
REFERENCES "per_game_stats" ("id");
```

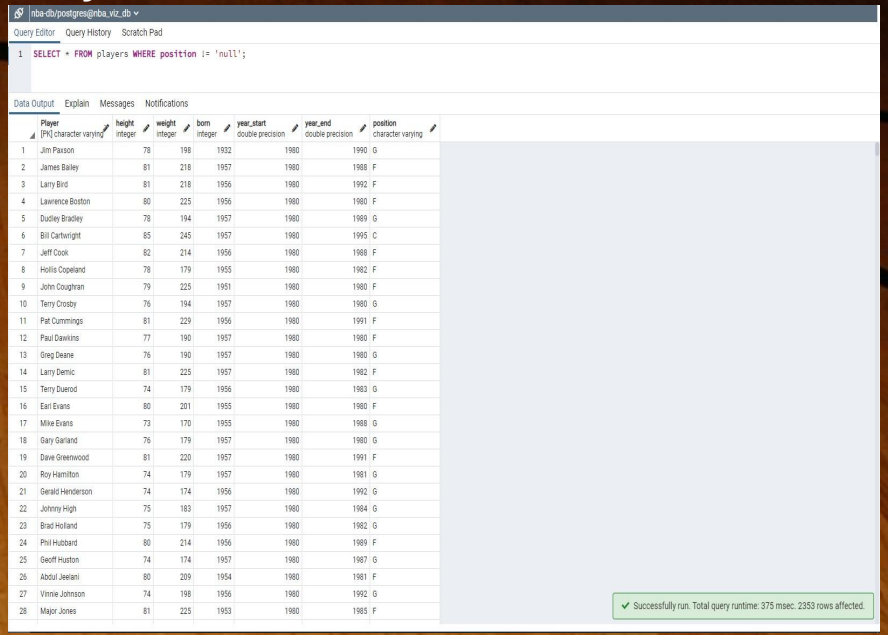
Database (cont'd)

Upload data into PostgreSQL database

Players data SQL table



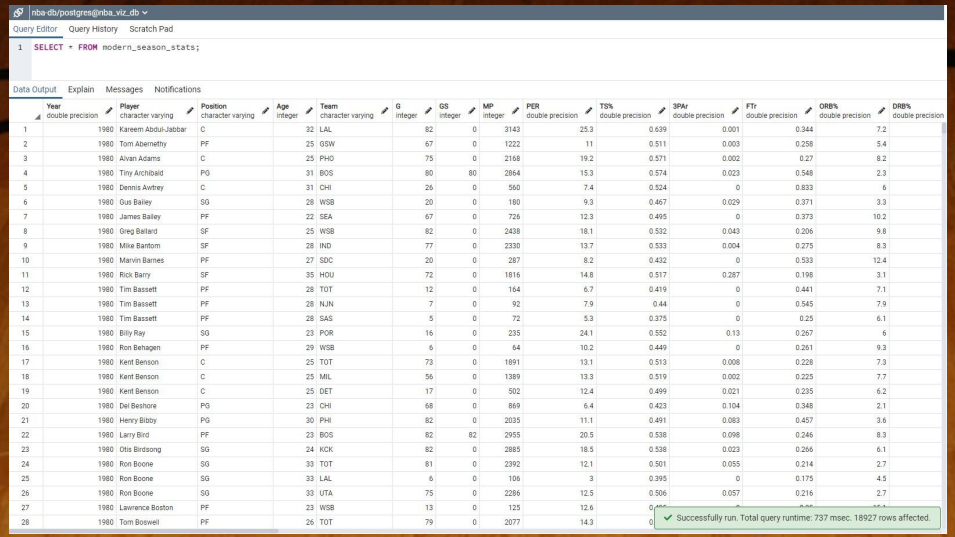
Players data SQL table without null values



Database (cont'd)

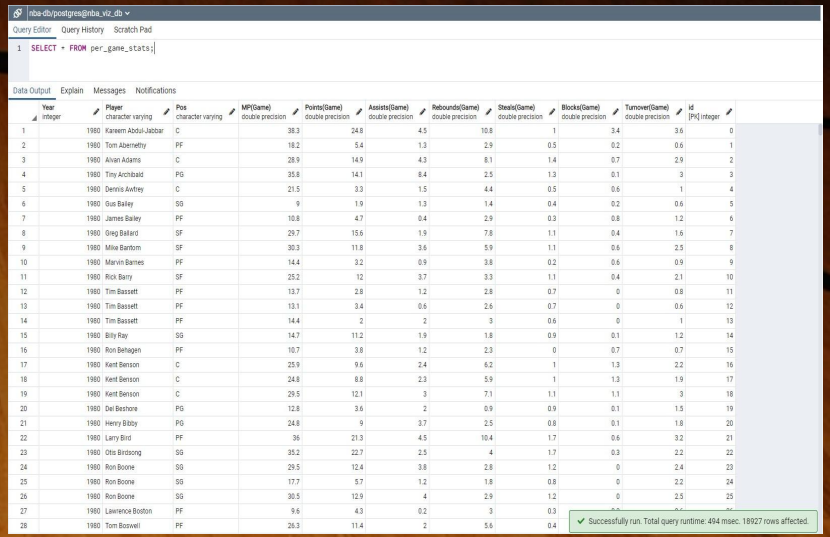
- Upload data into PostgreSQL database

modern season stats SQL table



	Year double precision	Player character varying	Position character varying	Age integer	Team character varying	G integer	GS integer	MP integer	PER double precision	TS% double precision	3PAr double precision	FTr double precision	ORB% double precision	DRB% double precision
1	1980	Kareem Abdul-Jabbar	C	32	LAL	82	0	3143	25.3	0.639	0.001	0.344	7.2	
2	1980	Tom Abernethy	PF	25	GSW	67	0	1222	11	0.511	0.003	0.258	5.4	
3	1980	Alvan Adams	C	25	PHO	75	0	2168	19.2	0.571	0.002	0.27	8.2	
4	1980	Tiny Archibald	PG	31	BOS	80	80	2864	15.3	0.574	0.023	0.548	2.3	
5	1980	Dennis Awtrey	C	31	CHI	26	0	560	7.4	0.524	0	0.833	6	
6	1980	Gus Bailey	SG	28	WSB	20	0	180	9.3	0.467	0.029	0.371	3.3	
7	1980	James Bailey	PF	22	SEA	67	0	726	12.3	0.495	0	0.373	10.2	
8	1980	Greg Ballard	SF	25	WSB	82	0	2438	18.1	0.532	0.043	0.206	9.8	
9	1980	Mike Bantom	SF	28	IND	77	0	2330	13.7	0.533	0.004	0.275	8.3	
10	1980	Marvin Barnes	PF	27	SDC	20	0	287	8.2	0.432	0	0.533	12.4	
11	1980	Rick Barry	SF	35	HOU	72	0	1816	14.8	0.517	0.287	0.198	3.1	
12	1980	Tim Bassett	PF	28	TOT	12	0	164	6.7	0.419	0	0.441	7.1	
13	1980	Tim Bassett	PF	28	NJN	7	0	92	7.9	0.44	0	0.545	7.9	
14	1980	Tim Bassett	PF	28	SAS	5	0	72	5.3	0.375	0	0.25	6.1	
15	1980	Billy Ray	SG	23	POR	16	0	235	24.1	0.552	0.13	0.267	6	
16	1980	Ron Behagen	PF	29	WSB	6	0	64	10.2	0.449	0	0.261	9.3	
17	1980	Kent Benson	C	25	TOT	73	0	1891	13.1	0.513	0.008	0.228	7.3	
18	1980	Kent Benson	C	25	MIL	56	0	1389	13.3	0.519	0.002	0.225	7.7	
19	1980	Kent Benson	C	25	DET	17	0	502	12.4	0.499	0.021	0.235	6.2	
20	1980	Del Beshore	PG	23	CHI	68	0	869	6.4	0.423	0.104	0.348	2.1	
21	1980	Henry Bibby	PG	30	PHI	82	0	2035	11.1	0.491	0.083	0.457	3.6	
22	1980	Larry Bird	PF	23	BOS	82	82	2955	20.5	0.538	0.098	0.246	8.3	
23	1980	Otis Birdsong	SG	24	KCK	82	0	2885	18.5	0.538	0.023	0.266	6.1	
24	1980	Ron Boone	SG	33	TOT	81	0	2392	12.1	0.501	0.055	0.214	2.7	
25	1980	Ron Boone	SG	33	LAL	6	0	106	3	0.395	0	0.175	4.5	
26	1980	Ron Boone	SG	33	UTA	75	0	2286	12.5	0.506	0.057	0.216	2.7	
27	1980	Lawrence Boston	PF	23	WSB	13	0	125	12.6	0.4				
28	1980	Tom Boswell	PF	26	TOT	79	0	2077	14.3	0.5				

per_game_stats SQL table

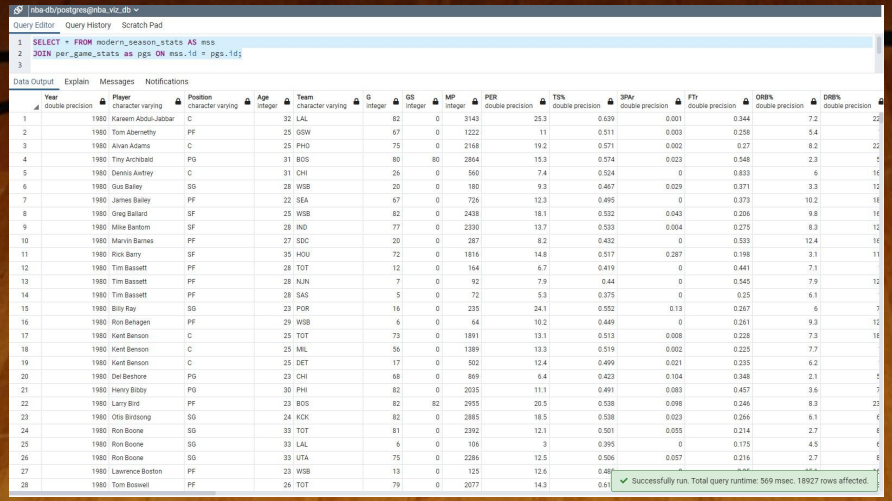


	Year integer	Player character varying	Pos character varying	MP(Game) double precision	Points(Game) double precision	Assists(Game) double precision	Rebounds(Game) double precision	Steals(Game) double precision	Blocks(Game) double precision	Turnover(Game) double precision	id [PK] integer
1	1980	Kareem Abdul-Jabbar	C	38.3	24.8	4.5	10.8	1	3.4	3.6	0
2	1980	Tom Abernethy	PF	18.2	5.4	1.3	2.9	0.5	0.2	0.6	1
3	1980	Alvan Adams	C	28.9	14.9	4.3	8.1	1.4	0.7	2.9	2
4	1980	Tiny Archibald	PG	35.8	14.1	8.4	2.5	1.3	0.1	3	3
5	1980	Dennis Awtrey	C	21.5	3.3	1.5	4.4	0.5	0.6	1	4
6	1980	Gus Bailey	SG	9	1.9	1.3	1.4	0.4	0.2	0.6	5
7	1980	James Bailey	PF	10.8	4.7	0.4	2.9	0.3	0.8	1.2	6
8	1980	Greg Ballard	SF	29.7	15.6	1.9	7.8	1.1	0.4	1.6	7
9	1980	Mike Bantom	SF	30.3	11.8	3.6	5.9	1.1	0.6	2.5	8
10	1980	Marvin Barnes	PF	14.4	3.2	0.9	3.8	0.2	0.6	0.9	9
11	1980	Rick Barry	SF	25.2	12	3.7	3.3	1.1	0.4	2.1	10
12	1980	Tim Bassett	PF	13.7	2.8	1.2	2.8	0.7	0	0.8	11
13	1980	Tim Bassett	PF	13.1	3.4	0.6	2.6	0.7	0	0.6	12
14	1980	Tim Bassett	PF	14.4	2	2	3	0.6	0	1	13
15	1980	Billy Ray	SG	14.7	11.2	1.9	1.8	0.9	0.1	1.2	14
16	1980	Ron Behagen	PF	10.7	3.8	1.2	2.3	0	0.7	0.7	15
17	1980	Kent Benson	C	25.9	9.6	2.4	6.2	1	1.3	2.2	16
18	1980	Kent Benson	C	24.8	8.8	2.3	5.9	1	1.3	1.9	17
19	1980	Kent Benson	C	29.5	12.1	3	7.1	1.1	1.1	3	18
20	1980	Del Beshore	PG	12.8	3.6	2	0.9	0.9	0.1	1.5	19
21	1980	Henry Bibby	PG	24.8	9	3.7	2.5	0.8	0.1	1.8	20
22	1980	Larry Bird	PF	36	21.3	4.5	10.4	1.7	0.6	3.2	21
23	1980	Otis Birdsong	SG	35.2	22.7	2.5	4	1.7	0.3	2.2	22
24	1980	Ron Boone	SG	29.5	12.4	3.8	2.8	1.2	0	2.4	23
25	1980	Ron Boone	SG	17.7	5.7	1.2	1.8	0.8	0	2.2	24
26	1980	Ron Boone	SG	30.5	12.9	4	2.9	1.2	0	2.5	25
27	1980	Lawrence Boston	PF	9.6	4.3	0.2	3	0.3			
28	1980	Tom Boswell	PF	26.3	11.4	2	5.6	0.4			

Database (cont'd)

- PostgreSQL database tables with joins

Joining modern season stats and per_game_stats SQL table

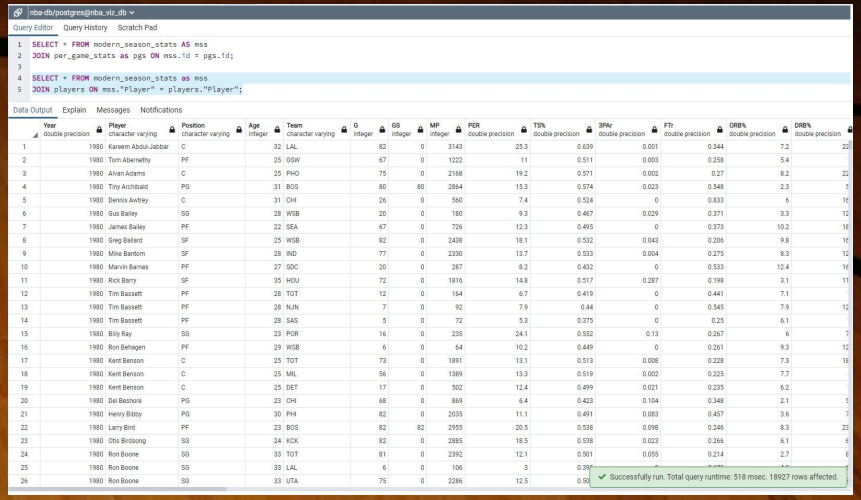


```
SELECT * FROM modern_season_stats AS mss
JOIN per_game_stats as pgs ON mss.id = pgs.id;
```

	Year double precision	Player character varying	Position character varying	Age integer	Team character varying	G integer	GS integer	MP integer	PER double precision	TS% double precision	3PAr double precision	FTr double precision	ORB% double precision	DRB% double precision
1	1980	Kareem Abdul-Jabbar	C	32	LAL	82	0	3143	25.3	0.639	0.001	0.344	7.2	22
2	1980	Tom Abernethy	PF	25	GSW	67	0	1222	11	0.511	0.003	0.258	5.4	
3	1980	Alvan Adams	C	25	PHO	75	0	2168	19.2	0.571	0.002	0.27	8.2	22
4	1980	Tiny Archibald	PG	31	BOS	80	80	2864	15.3	0.574	0.023	0.548	2.3	5
5	1980	Dennis Awtrey	C	31	CHI	26	0	560	7.4	0.524	0	0.833	6	16
6	1980	Gus Bailey	SG	28	WSB	20	0	180	9.3	0.467	0.029	0.371	3.3	12
7	1980	James Bailey	PF	22	SEA	67	0	726	12.3	0.495	0	0.373	10.2	16
8	1980	Greg Ballard	SF	25	WSB	82	0	2438	18.1	0.532	0.043	0.206	9.8	16
9	1980	Mike Bantom	SF	28	IND	77	0	2330	13.7	0.533	0.004	0.275	8.3	12
10	1980	Marvin Barnes	PF	27	SDC	20	0	287	8.2	0.432	0	0.533	12.4	16
11	1980	Rick Barry	SF	35	HOU	72	0	1816	14.8	0.517	0.287	0.198	3.1	11
12	1980	Tim Bassett	PF	28	TOT	12	0	164	6.7	0.419	0	0.441	7.1	
13	1980	Tim Bassett	PF	28	NJN	7	0	92	7.9	0.44	0	0.545	7.9	12
14	1980	Tim Bassett	PF	28	SAS	5	0	72	5.3	0.375	0	0.25	6.1	
15	1980	Billy Ray	SG	23	POR	16	0	235	24.1	0.552	0.13	0.267	6	7
16	1980	Ron Behagen	PF	29	WSB	6	0	64	10.2	0.449	0	0.261	9.3	12
17	1980	Kent Benson	C	25	TOT	73	0	1891	13.1	0.513	0.008	0.228	7.3	18
18	1980	Kent Benson	C	25	MIL	56	0	1389	13.3	0.519	0.002	0.225	7.7	
19	1980	Kent Benson	C	25	DET	17	0	502	12.4	0.499	0.021	0.235	6.2	
20	1980	Del Beshore	PG	23	CHI	68	0	869	6.4	0.423	0.104	0.348	2.1	5
21	1980	Henry Bibby	PG	30	PHI	82	0	2035	11.1	0.491	0.083	0.457	3.6	7
22	1980	Larry Bird	PF	23	BOS	82	82	2955	20.5	0.538	0.098	0.246	8.3	23
23	1980	Otis Birdsong	SG	24	KCK	82	0	2885	18.5	0.538	0.023	0.266	6.1	6
24	1980	Ron Boone	SG	33	TOT	81	0	2392	12.1	0.501	0.055	0.214	2.7	8
25	1980	Ron Boone	SG	33	LAL	6	0	106	3	0.395	0	0.175	4.5	6
26	1980	Ron Boone	SG	33	UTA	75	0	2286	12.5	0.506	0.057	0.216	2.7	8
27	1980	Lawrence Boston	PF	23	WSB	13	0	125	12.6	0.48				
28	1980	Tom Boswell	PF	26	TOT	79	0	2077	14.3	0.61				

Successfully run. Total query runtime: 569 msec. 18927 rows affected.

Joining modern_season_stats and players SQL table



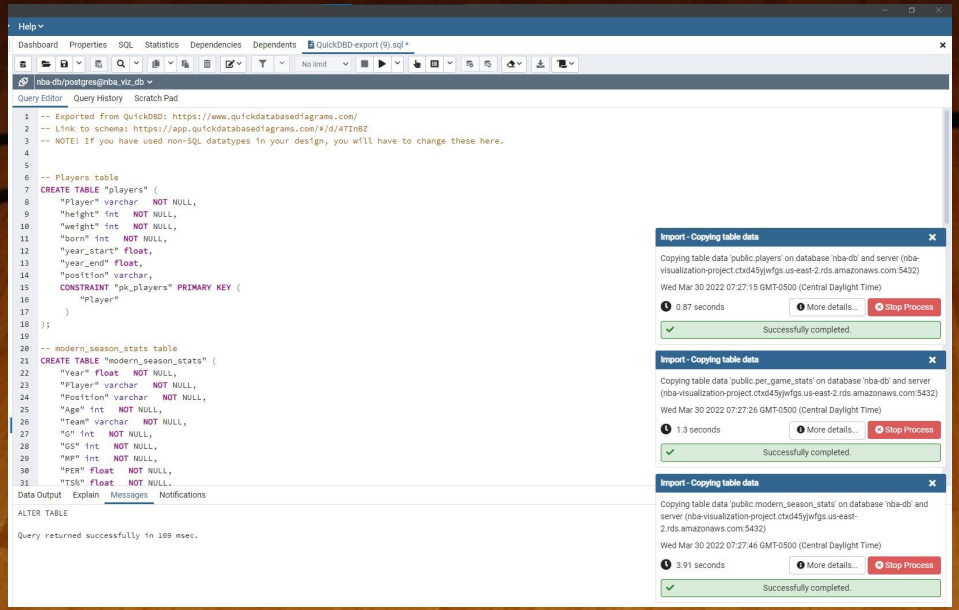
```
SELECT * FROM modern_season_stats AS mss
JOIN per_game_stats as pgs ON mss.id = pgs.id;

SELECT * FROM modern_season_stats as mss
JOIN players ON mss."Player" = players."Player";
```

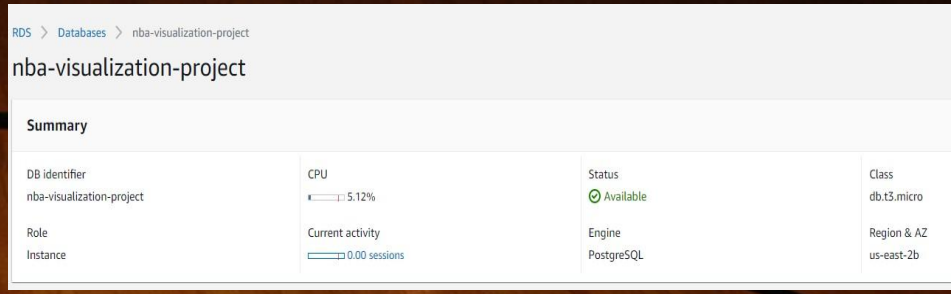
	Year double precision	Player character varying	Position character varying	Age integer	Team character varying	G integer	GS integer	MP integer	PER double precision	TS% double precision	3PAr double precision	FTr double precision	ORB% double precision	DRB% double precision
1	1980	Kareem Abdul-Jabbar	C	32	LAL	82	0	3143	25.3	0.639	0.001	0.344	7.2	22
2	1980	Tom Abernethy	PF	25	GSW	67	0	1222	11	0.511	0.003	0.258	5.4	
3	1980	Alvan Adams	C	25	PHO	75	0	2168	19.2	0.571	0.002	0.27	8.2	22
4	1980	Tiny Archibald	PG	31	BOS	80	80	2864	15.3	0.574	0.023	0.548	2.3	5
5	1980	Dennis Awtrey	C	31	CHI	26	0	560	7.4	0.524	0	0.833	6	16
6	1980	Gus Bailey	SG	28	WSB	20	0	180	9.3	0.467	0.029	0.371	3.3	12
7	1980	James Bailey	PF	22	SEA	67	0	726	12.3	0.495	0	0.373	10.2	16
8	1980	Greg Ballard	SF	25	WSB	82	0	2438	18.1	0.532	0.043	0.206	9.8	16
9	1980	Mike Bantom	SF	28	IND	77	0	2330	13.7	0.533	0.004	0.275	8.3	12
10	1980	Marvin Barnes	PF	27	SDC	20	0	287	8.2	0.432	0	0.533	12.4	16
11	1980	Rick Barry	SF	35	HOU	72	0	1816	14.8	0.517	0.287	0.198	3.1	11
12	1980	Tim Bassett	PF	28	TOT	12	0	164	6.7	0.419	0	0.441	7.1	
13	1980	Tim Bassett	PF	28	NJN	7	0	92	7.9	0.44	0	0.545	7.9	12
14	1980	Tim Bassett	PF	28	SAS	5	0	72	5.3	0.375	0	0.25	6.1	
15	1980	Billy Ray	SG	23	POR	16	0	235	24.1	0.552	0.13	0.267	6	7
16	1980	Ron Behagen	PF	29	WSB	6	0	64	10.2	0.449	0	0.261	9.3	12
17	1980	Kent Benson	C	25	TOT	73	0	1891	13.1	0.513	0.008	0.228	7.3	18
18	1980	Kent Benson	C	25	MIL	56	0	1389	13.3	0.519	0.002	0.225	7.7	
19	1980	Kent Benson	C	25	DET	17	0	502	12.4	0.499	0.021	0.235	6.2	
20	1980	Del Beshore	PG	23	CHI	68	0	869	6.4	0.423	0.104	0.348	2.1	5
21	1980	Henry Bibby	PG	30	PHI	82	0	2035	11.1	0.491	0.083	0.457	3.6	7
22	1980	Larry Bird	PF	23	BOS	82	82	2955	20.5	0.538	0.098	0.246	8.3	23
23	1980	Otis Birdsong	SG	24	KCK	82	0	2885	18.5	0.538	0.023	0.266	6.1	6
24	1980	Ron Boone	SG	33	TOT	81	0	2392	12.1	0.501	0.055	0.214	2.7	8
25	1980	Ron Boone	SG	33	LAL	6	0	106	3	0.395				
26	1980	Ron Boone	SG	33	UTA	75	0	2286	12.5	0.506				

Successfully run. Total query runtime: 518 msec. 18927 rows affected.

Database (cont'd)

- Create AWS RDS and connect to PostgreSQL





Database created in AWS

PostgreSQL database connected to AWS

Database (cont'd)

- Connect AWS RDS to Python with SQLAlchemy

```
[ ]  1 import sqlalchemy
     2 from sqlalchemy.ext.automap import automap_base
     3 from sqlalchemy.orm import Session
     4 from sqlalchemy import create_engine, func
     5 import psycopg2
     6 import pandas as pd
     7 from psycopg2 import sql
```

```
[ ]  1 # SQLAlchemy create_engine('postgresql://username:password@host/db_name)
     2 engine = create_engine('postgresql://postgres[redacted]visualization-project.ctxd45yjwfgs.us-east-2.rds.amazonaws.com/nba-db')
```

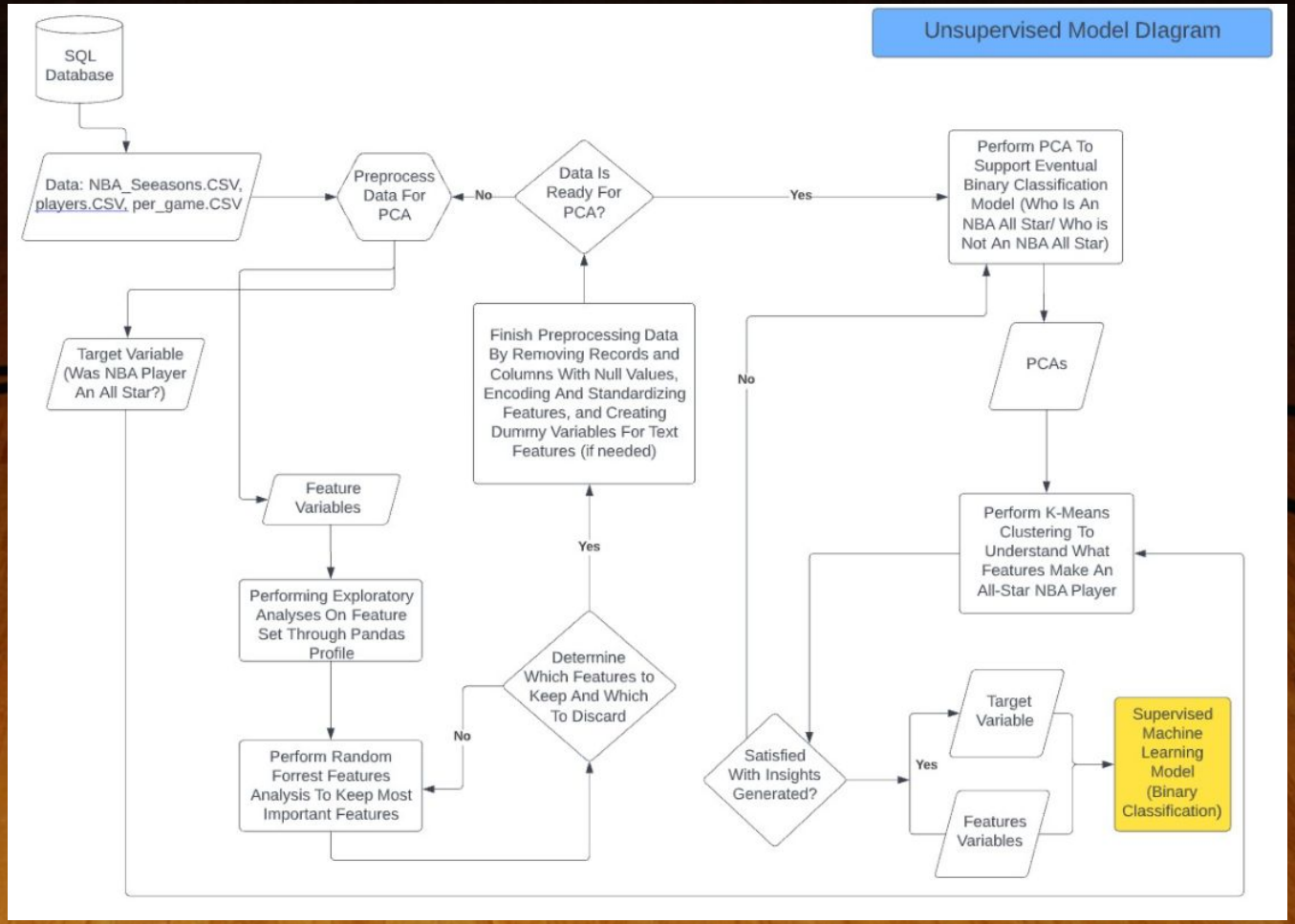
```
[ ]  1 Base = automap_base()
     2 Base.prepare(engine, reflect=True)
```

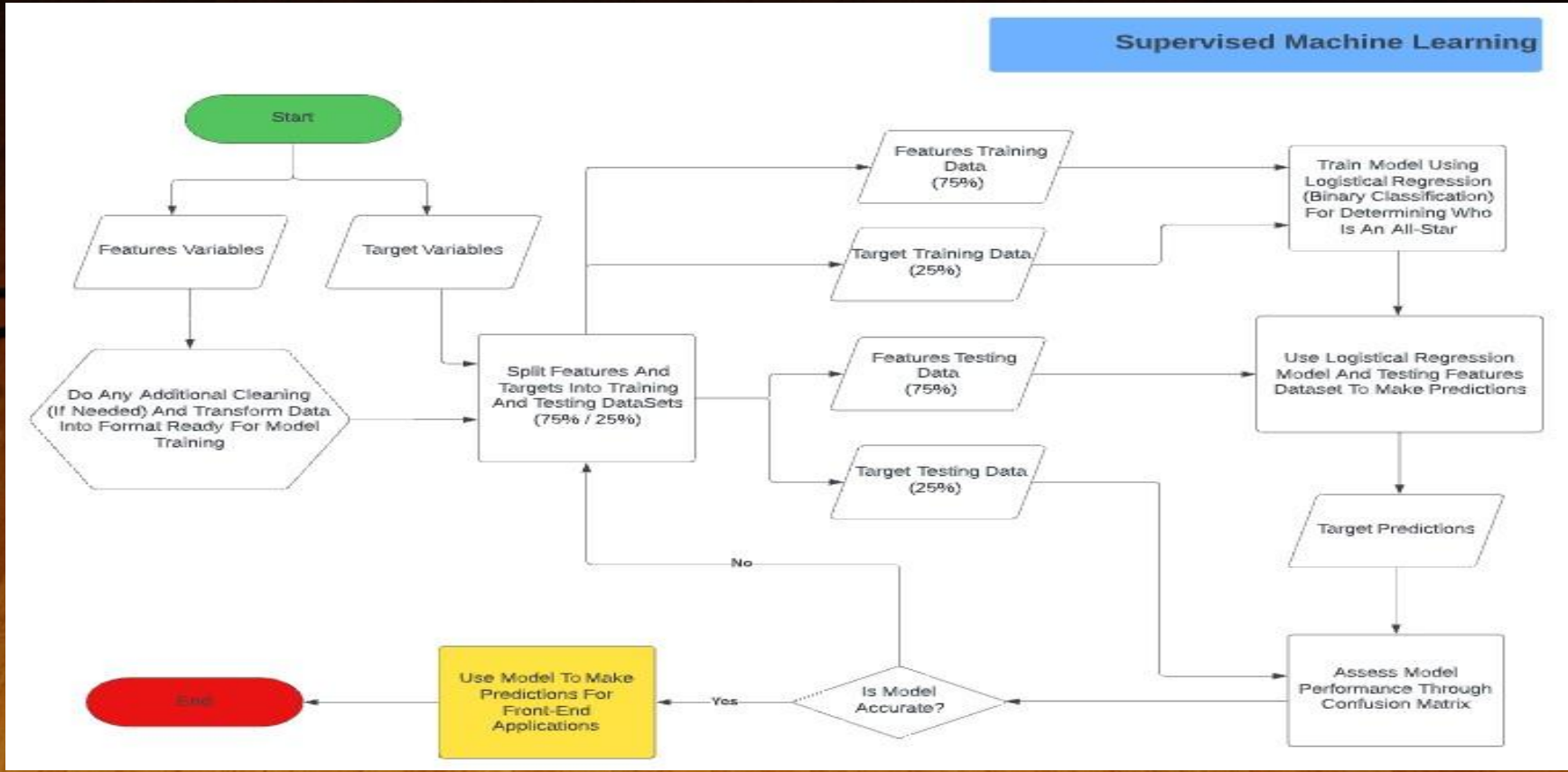
```
[ ]  1 # query syntax: df = pd.read_sql(query.statement, connection)
     2
     3 ## players DF from players table
     4 players_df = pd.read_sql('SELECT * FROM players', engine)
     5 ## seasons_df from modern_season_stats table
     6 seasons_df = pd.read_sql('SELECT * FROM modern_season_stats', engine)
     7 ## per_game_df  from per_game_stats table
     8 per_game_df = pd.read_sql('SELECT * FROM per_game_stats', engine)
```

```
[ ]  1 per_game_df.where(per_game_df.Player == 'Michael Jordan').dropna()
```

	Year	Player	Pos	MP(Game)	Points(Game)	Assists(Game)	Rebounds(Game)	Steals(Game)	Blocks(Game)	Turnover(Game)	id
1984	1985.0	Michael Jordan	SG	38.3	28.2	5.9	6.5	2.4	0.8	3.5	1984.0
2361	1986.0	Michael Jordan	SG	25.1	22.7	2.9	3.6	2.1	1.2	2.5	2361.0
2735	1987.0	Michael Jordan	SG	40.0	37.1	4.6	5.2	2.9	1.5	3.3	2735.0
3134	1988.0	Michael Jordan	SG	40.4	35.0	5.9	5.5	3.2	1.6	3.1	3134.0
3561	1989.0	Michael Jordan	SG	40.2	32.5	8.0	8.0	2.9	0.8	3.6	3561.0
4010	1990.0	Michael Jordan	SG	39.0	33.6	6.3	6.9	2.8	0.7	3.0	4010.0

Machine Learning

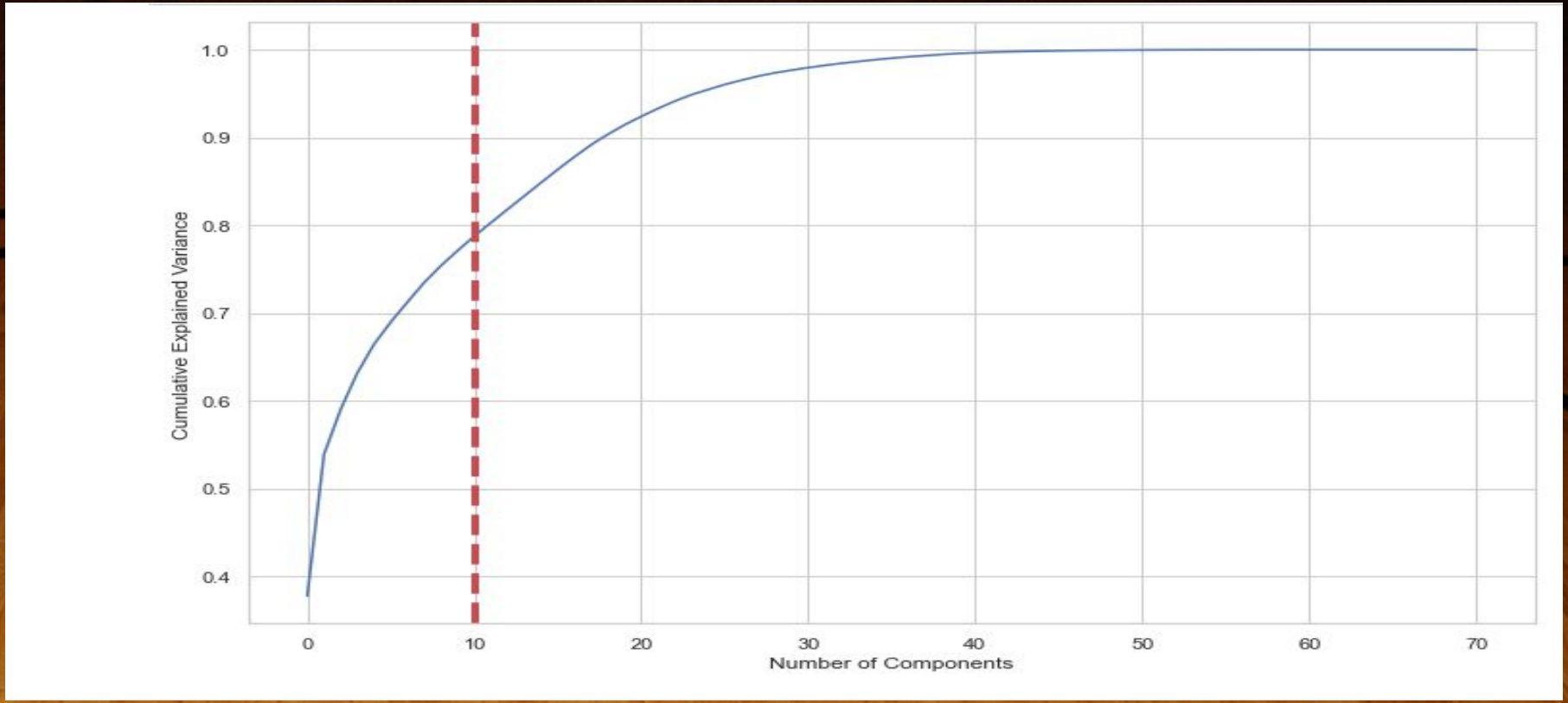

Unsupervised Model DIagram
SQL Database
Data: NBA_Seeasons.CSV, players.CSV, per_game.CSV
Preprocess Data For PCA
No
Data Is Ready For PCA?
Yes
Perform PCA To Support Eventual Binary Classification Model (Who Is An NBA All Star/ Who is Not An NBA All Star)
Target Variable (Was NBA Player An All Star?)
Finish Preprocessing Data By Removing Records and Columns With Null Values, Encoding And Standardizing Features, and Creating Dummy Variables For Text Features (if needed)
No
PCAs
Feature Variables
Yes
Perform K-Means Clustering To Understand What Features Make An All-Star NBA Player
Performing Exploratory Analyses On Feature Set Through Pandas Profile
Determine Which Features to Keep And Which To Discard
No
Perform Random Forrest Features Analysis To Keep Most Important Features
Satisfied With Insights Generated?
Yes
Target Variable
Features Variables
Supervised Machine Learning Model (Binary Classification)


Supervised Machine Learning
Start
Features Variables
Target Variables
Do Any Additional Cleaning (If Needed) And Transform Data Into Format Ready For Model Training
Split Features And Targets Into Training And Testing DataSets (75% / 25%)
Features Training Data (75%)
Target Training Data (25%)
Features Testing Data (75%)
Target Testing Data (25%)
Train Model Using Logistical Regression (Binary Classification) For Determining Who Is An All-Star
Use Logistical Regression Model And Testing Features Dataset To Make Predictions
Target Predictions
Assess Model Performance Through Confusion Matrix
Is Model Accurate?
No
Yes
Use Model To Make Predictions For Front-End Applications
End

Pre-Processing

- Games ⇒ NBA AllStar
- Initially ranked by feature importance through random forests without PCA.
- Saw overrepresentation issues with non-NBA AllStar players so performed over- and under-sampling.
- Decided to go with under-sampling and applied it as a pre-processing step for ALL OF OUR MODELS.
- Performed other cleaning tasks for the features.
- Target was whether player would be an NBA AllStar (1) or not (0).

PCA: Going From 68 to 10!

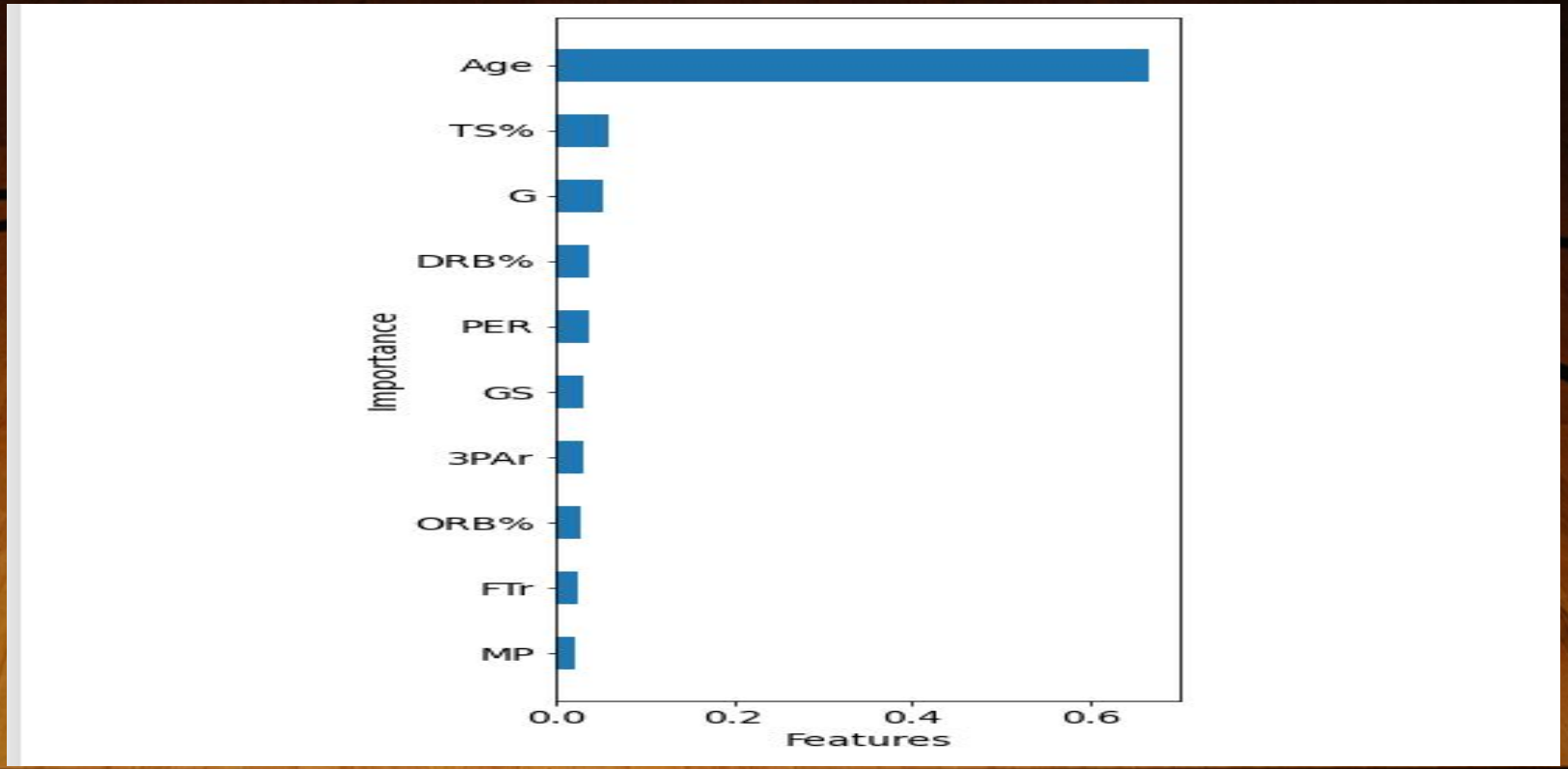


PCA: Top Ten Features Explanatory Power

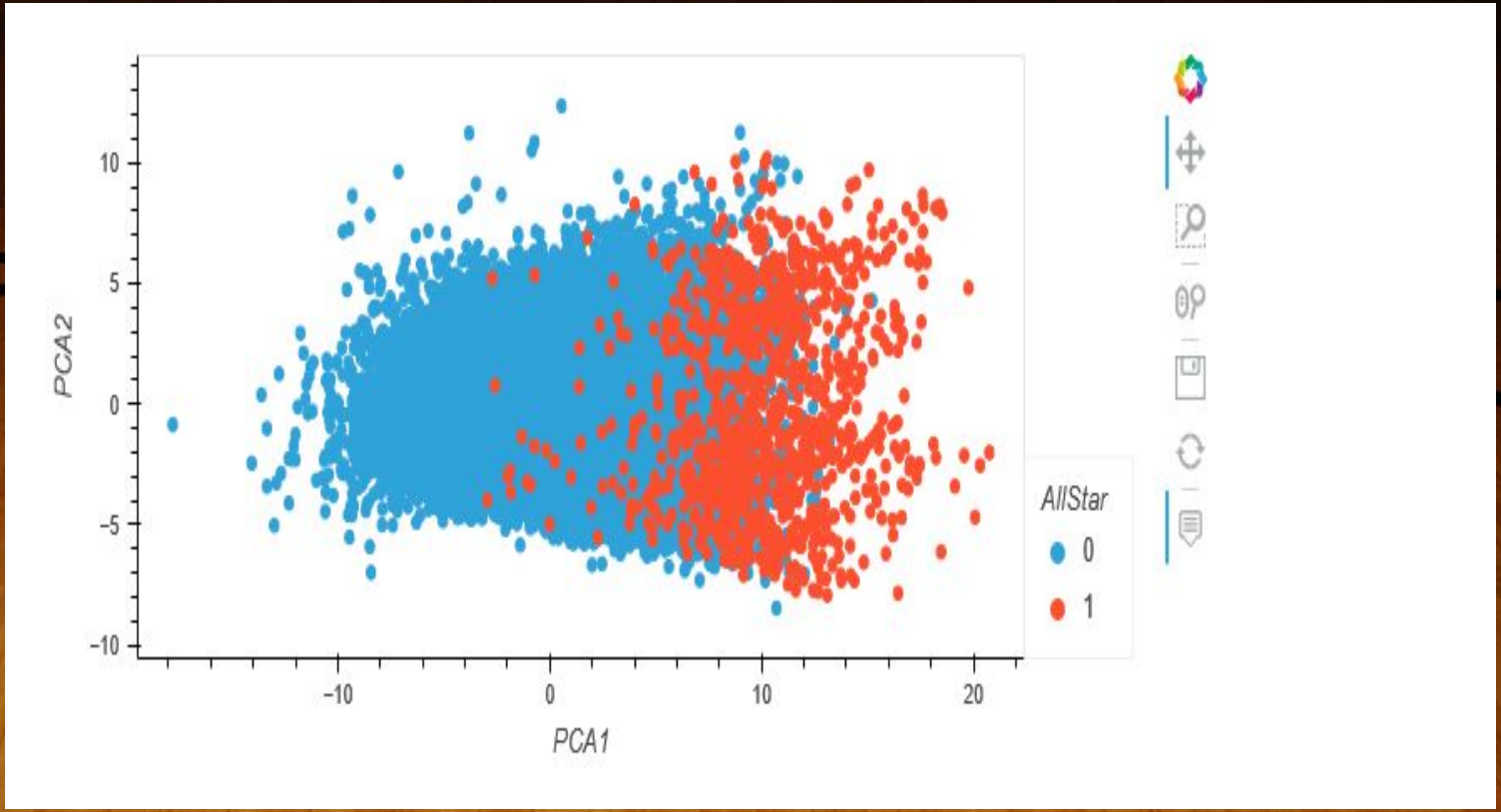
Out[25]:

	Cumulative Variance Ratio	Explained Variance Ratio
0	0.377207	0.377207
1	0.538460	0.161253
2	0.589985	0.051525
3	0.631495	0.041511
4	0.664259	0.032763
5	0.689535	0.025276
6	0.712278	0.022743
7	0.734474	0.022196
8	0.753694	0.019219
9	0.770883	0.017190

PCA: Top Ten Features



K-Means



Supervised Models

- Metrics of Interests:
  - Accuracy
  - Recall
  - F-1 score
- Binary Classification!!!
- Models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - Random Forests (with top ten variables this time!)
  - Boosting

Summary of Supervised Performances!

Metric	Logistical Regression	SVM	Decision Tree	Random Forest	Boosting
Testing Accuracy	91.1%	90.7%	89.3%	92.4%	91.9%
Recall (for predicting non-AllStars)	91%	91%	90%	92%	92%
Recall (for predicting AllStars)	94%	94%	85%	94%	92%
F-1 Scores (for predicting AllStars)	51%	50%	44%	55%	53%

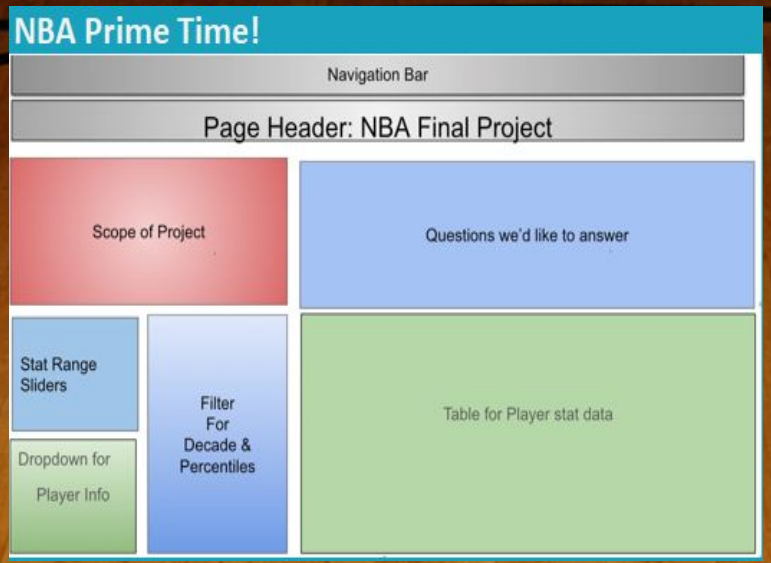
Dashboard Storyboards & Mockups

Storyboards & Dashboard

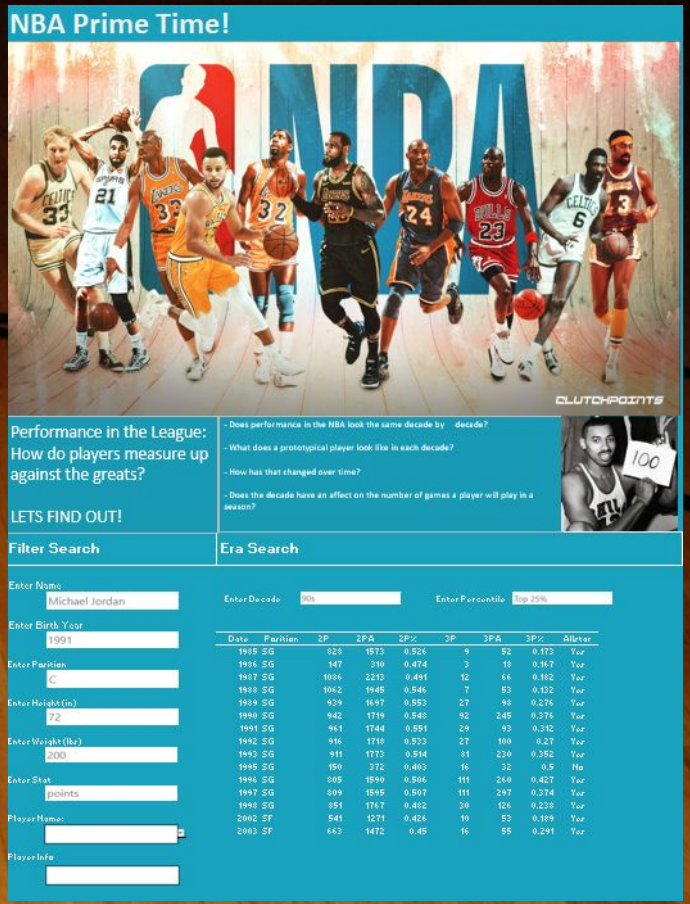
- A blueprint for the dashboard is created and includes all of the following:
  - Storyboard on Google Slide(s)
  - Description of tool(s) that will be used to create final dashboard
  - Description of interactive element(s)
- Tools used :
  - HTML
  - CSS
  - JavaScript
  - GitHub Pages

Dashboard - Page 1

Storyboard Layout

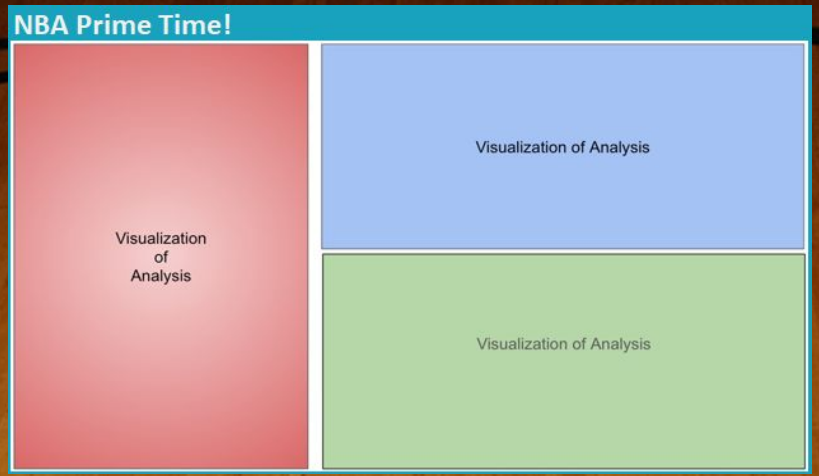


Mockup Page

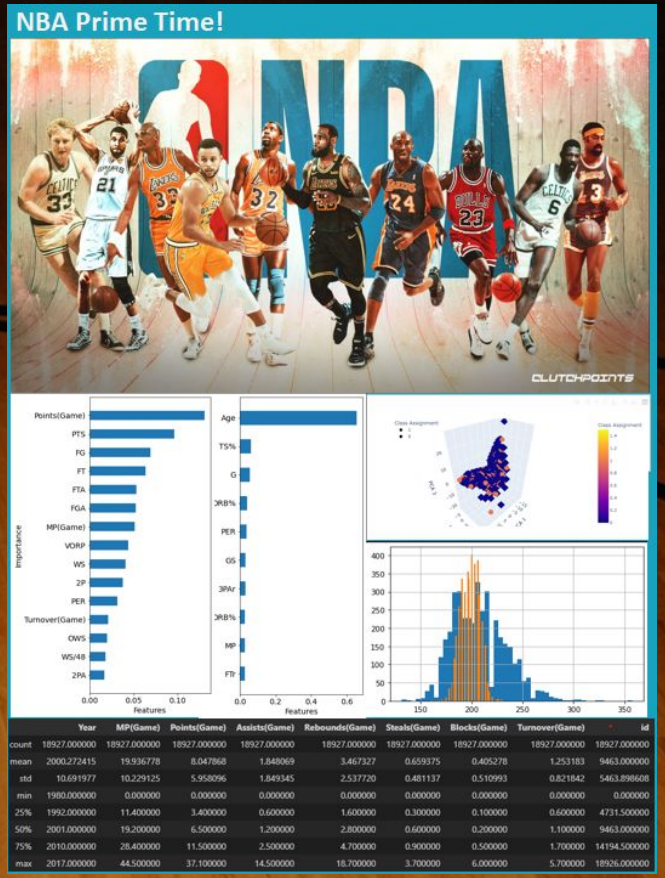


Dashboard - Page 2

Storyboard Layout

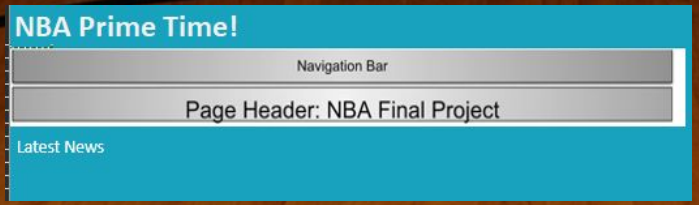


Mockup Page



Dashboard - Page 3

Storyboard Layout



Mockup Page