6.1: Sourcing Open Data

Data Source -

The data is an external dataset, and the dataset was pulled from Kaggle, The players data for Career Mode from FIFA 20 ("players 20.csv") is included in the datasets. The information enables for analysis of players and their different attributes.

Data Collection -

Every year FIFA releases different version of games, and the dataset was made public by online gaming platform. Link - https://sofifa.com/ and published on Kaggle for personal use.

An explanation for why you've chosen this data set -

I'm interested in this information because I've always been fascinated by sports. I'm a gamer myself, and I'd like to put my talents to the test by analyzing the players and their attributes in this game version. Three years ago, the FIFA dataset was updated.

Data Profiling -

The original dataset contains 18278 rows and 104 columns, but 64 columns were dropped because they weren't needed for this analysis.

Original data column below -

data.csv includes lastest edition FIFA 2020 players attributes like Age, Nationality, Overall, Potential, Club, Value, Wage, Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Jersey Number, Joined, Loaned From, Contract Valid Until, Height, Weight, LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB, Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes, and Release Clause.

Below is the column used for this analysis -

sofifa_id, short_name, long_name, age, height_cm, weight_kg, nationality, club, overall, potential, value_eur, wage_eur, preferred_foot, international_reputation, weak_foot, skill_moves, work_rate, body_type, release_clause_eur, team_position, joined, contract_valid_until, p *ace, shooting , passing, dribbling, defending, physic, player_traits, power_shot_power, power_jumping, power_long_shots, mentality_aggression, mentality_interceptions, mentality_penalties.*

Consistency Checks & Cleaning -

Data Types:

There were two mixed data types, and both were change to strings.
(team_position and player_traits)

Joined converted to string

Body_type column was cleaned by changing the variables below*-*

Change Messi to lean
Change C. Ronaldo to normal
Change Neymar to lean
Change PLAYER_BODY_TYPE_25 to normal
Change Courtois to lean
Change Shaqiri to stocky
Change Akinfenwa to stocky

Check for uniqueness*-*
18038 rows are unique

Missing Values -

- *contract_valid_until column has few missing values,*
- *I created a subset of the data frame containing only those values within the "contract_valid_until" column that meet the missing value condition.*
- *while I impute missing values with mean in the numerical values.*
- *creating a new data frame and filter out the ones that aren't missing into a subset data frame and continue with the analysis with this new data frame.*

Dropped Columns after cleaning

- *player_traits column – wasn't needed for this analysis*
- *After cleaning the fifa20 dataset, 18038 rows and 34 columns remaining for this analysis*

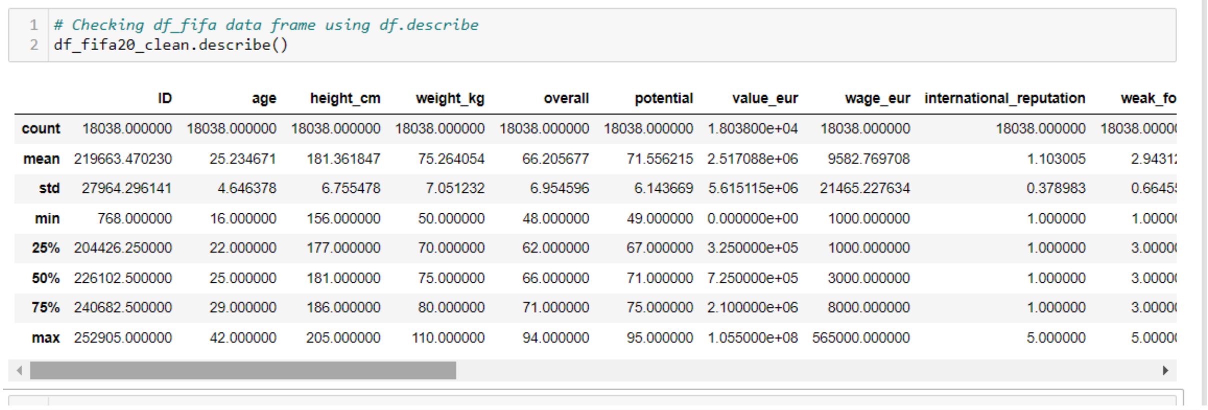
Duplicate
No duplicate found

Renaming-

Rename sofifa_id to ID

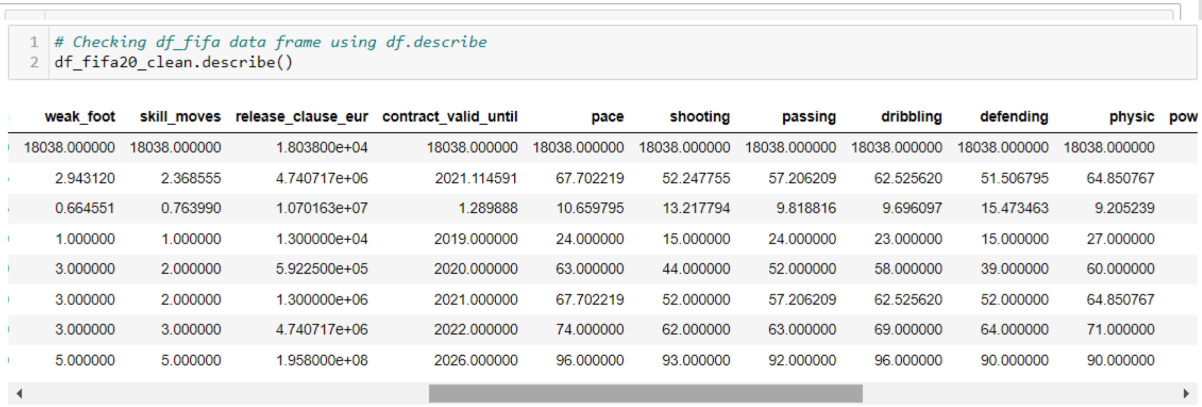
Understand your data. Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis. Summary statistics-

Descriptive statistics -



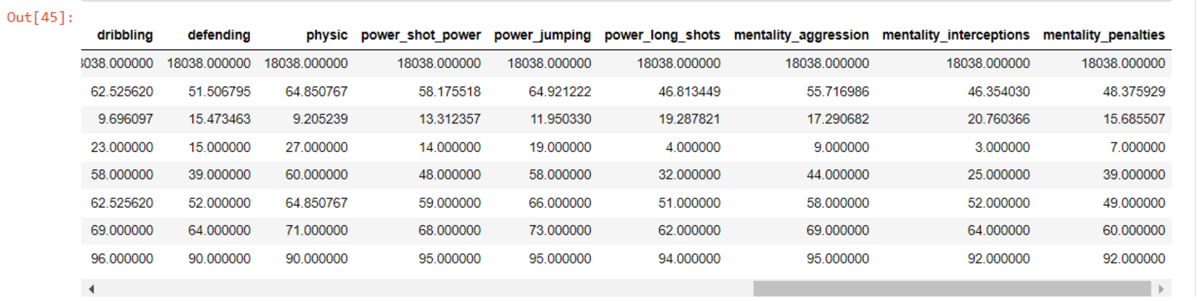
```
# Checking df_fifa data frame using df.describe
df_fifa20_clean.describe()
```

	ID	age	height_cm	weight_kg	overall	potential	value_eur	wage_eur	international_reputation	weak_fo
count	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	1.803800e+04	18038.000000	18038.000000	18038.0000
mean	219663.470230	25.234671	181.361847	75.264054	66.205677	71.556215	2.517088e+06	9582.769708	1.103005	2.9431
std	27964.296141	4.646378	6.755478	7.051232	6.954596	6.143669	5.615115e+06	21465.227634	0.378983	0.6645
min	768.000000	16.000000	156.000000	50.000000	48.000000	49.000000	0.000000e+00	1000.000000	1.000000	1.0000
25%	204426.250000	22.000000	177.000000	70.000000	62.000000	67.000000	3.250000e+05	1000.000000	1.000000	3.0000
50%	226102.500000	25.000000	181.000000	75.000000	66.000000	71.000000	7.250000e+05	3000.000000	1.000000	3.0000
75%	240682.500000	29.000000	186.000000	80.000000	71.000000	75.000000	2.100000e+06	8000.000000	1.000000	3.0000
max	252905.000000	42.000000	205.000000	110.000000	94.000000	95.000000	1.055000e+08	565000.000000	5.000000	5.0000



```
# Checking df_fifa data frame using df.describe
df_fifa20_clean.describe()
```

weak_foot	skill_moves	release_clause_eur	contract_valid_until	pace	shooting	passing	dribbling	defending	physic	pow
18038.000000	18038.000000	1.803800e+04	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	
2.943120	2.368555	4.740717e+06	2021.114591	67.702219	52.247755	57.206209	62.525620	51.506795	64.850767	
0.664551	0.763990	1.070163e+07	1.289888	10.659795	13.217794	9.818816	9.696097	15.473463	9.205239	
1.000000	1.000000	1.300000e+04	2019.000000	24.000000	15.000000	24.000000	23.000000	15.000000	27.000000	
3.000000	2.000000	5.922500e+05	2020.000000	63.000000	44.000000	52.000000	58.000000	39.000000	60.000000	
3.000000	2.000000	1.300000e+06	2021.000000	67.702219	52.000000	57.206209	62.525620	52.000000	64.850767	
3.000000	3.000000	4.740717e+06	2022.000000	74.000000	62.000000	63.000000	69.000000	64.000000	71.000000	
5.000000	5.000000	1.958000e+08	2026.000000	96.000000	93.000000	92.000000	96.000000	90.000000	90.000000	



Out[45]:

dribbling	defending	physic	power_shot_power	power_jumping	power_long_shots	mentality_aggression	mentality_interceptions	mentality_penalties
038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000	18038.000000
62.525620	51.506795	64.850767	58.175518	64.921222	46.813449	55.716986	46.354030	48.375929
9.696097	15.473463	9.205239	13.312357	11.950330	19.287821	17.290682	20.760366	15.685507
23.000000	15.000000	27.000000	14.000000	19.000000	4.000000	9.000000	3.000000	7.000000
58.000000	39.000000	60.000000	48.000000	58.000000	32.000000	44.000000	25.000000	39.000000
62.525620	52.000000	64.850767	59.000000	66.000000	51.000000	58.000000	52.000000	49.000000
69.000000	64.000000	71.000000	68.000000	73.000000	62.000000	69.000000	64.000000	60.000000
96.000000	90.000000	90.000000	95.000000	95.000000	94.000000	95.000000	92.000000	92.000000

Consider limitations and ethics

There are no ethical concerns, no personal data in the dataset because anybody that plays FIFA games has access to their formation and so this information are made free to the public.

Define questions to explore. In a third section of your project document, define a list of questions to explore with your analysis-

- *Which age has the highest price value?*
- *Which player has the highest wage?*
- *Players with high wage what country are they from?*
- *Which nationals has younger players?*
- *Which club pays the highest wage?*
- *Which club has the tallest player?*
- *Is there a link between age, overall rating, and wage?*
- *Does body type have an impact on overall rating and wage?*
- *Is there a link between age and their potential?*
- *Does Nationality affect overall rating and price value?*
- *Does Nationality affect overall rating and wage*
- *Does Height affect, overall rating, and price value?*
- *Does Weight affect, overall rating, and price value?*