



## ▼ FIFA Football Analytics



### DATASET OVERVIEW

```
[2]: # Loading the required liabries.  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns
```

```
[5]: # Import the fifa dataset.  
data = pd.read_csv('players_20.csv')  
df = pd.DataFrame(data)  
df
```

```
[5]:
```

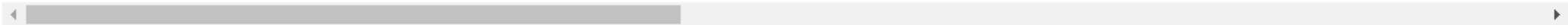
```
[3]: # Looking at the number of rows and columns in the dataset.  
df.shape
```

```
[3]: (18278, 104)
```

[4]: `# Basic stats about the dataset.`  
`df.describe()`

	sofifa_id	age	height_cm	weight_kg	overall	potential	value_eur	wage_eur	international_reputation	weak_foot	...	n
count	18278.000000	18278.000000	18278.000000	18278.000000	18278.000000	18278.000000	1.827800e+04	18278.000000	18278.000000	18278.000000	...	
mean	219738.864482	25.283291	181.362184	75.276343	66.244994	71.546887	2.484038e+06	9456.942773	1.103184	2.944250	...	
std	27960.200461	4.656964	6.756961	7.047744	6.949953	6.139669	5.585481e+06	21351.714095	0.378861	0.664656	...	
min	768.000000	16.000000	156.000000	50.000000	48.000000	49.000000	0.000000e+00	0.000000	1.000000	1.000000	...	
25%	204445.500000	22.000000	177.000000	70.000000	62.000000	67.000000	3.250000e+05	1000.000000	1.000000	3.000000	...	
50%	226165.000000	25.000000	181.000000	75.000000	66.000000	71.000000	7.000000e+05	3000.000000	1.000000	3.000000	...	
75%	240795.750000	29.000000	186.000000	80.000000	71.000000	75.000000	2.100000e+06	8000.000000	1.000000	3.000000	...	
max	252905.000000	42.000000	205.000000	110.000000	94.000000	95.000000	1.055000e+08	565000.000000	5.000000	5.000000	...	

8 rows × 61 columns



[5]: `# The columns.`  
`df.info()`

[5]: *# The columns.*

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 18278 entries, 0 to 18277  
Columns: 104 entries, sofifa_id to rb  
dtypes: float64(16), int64(45), object(43)  
memory usage: 14.5+ MB
```

[6]: *# Name of all the columns.*

```
for i in df.columns:  
    print(i)
```

```
sofifa_id  
player_url  
short_name  
long_name  
age  
dob  
height_cm  
weight_kg  
nationality  
club  
overall  
potential  
value_eur  
wage_eur
```

```
[7]: # Top 5 records of the dataset.  
df.head()
```

	sofifa_id	player_url	short_name	long_name	age	dob	height_cm	weight_kg	nationality	club	...	lwb	ldm	cdm	rdm	rv
0	158023	<a href="https://sofifa.com/player/158023/lionel-messi/...">https://sofifa.com/player/158023/lionel-messi/...</a>	L. Messi	Lionel Andrés Messi Cuccittini	32	1987-06-24	170	72	Argentina	FC Barcelona	...	68+2	66+2	66+2	66+2	68
1	20801	<a href="https://sofifa.com/player/20801/c-ronaldo-dos-...">https://sofifa.com/player/20801/c-ronaldo-dos-...</a>	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	34	1985-02-05	187	83	Portugal	Juventus	...	65+3	61+3	61+3	61+3	65
2	190871	<a href="https://sofifa.com/player/190871/neymar-da-sil...">https://sofifa.com/player/190871/neymar-da-sil...</a>	Neymar Jr	Neymar da Silva Santos Junior	27	1992-02-05	175	68	Brazil	Paris Saint- Germain	...	66+3	61+3	61+3	61+3	66
3	200389	<a href="https://sofifa.com/player/200389/jan-oblak/20/...">https://sofifa.com/player/200389/jan-oblak/20/...</a>	J. Oblak	Jan Oblak	26	1993-01-07	188	87	Slovenia	Atlético Madrid	...	NaN	NaN	NaN	NaN	N.
4	183277	<a href="https://sofifa.com/player/183277/eden-hazard/2...">https://sofifa.com/player/183277/eden-hazard/2...</a>	E. Hazard	Eden Hazard	28	1991-01-07	175	74	Belgium	Real Madrid	...	66+3	63+3	63+3	63+3	66

5 rows × 104 columns

## DATA CLEANING

```
[8]: # Extract required columns from the main dataset and make it another dataset named fifa.

fifa=df[['short_name','age','dob','height_cm','weight_kg','nationality','club','wage_eur','preferred_foot','international_reputation','pace','shooting','
fifa
```

[illegible]

```
# Check all the duplicate values in dataset.
fifa.duplicated().sum()
```

```
np.int64(0)
```

```
# Check all the null values in dataset.
fifa.isnull()
```

[illegible]

```
[11]: # Sum of all the null values in dataset.  
fifa.isnull().sum()
```

```
[11]: short_name          0  
age                    0  
dob                   0  
height_cm             0  
weight_kg             0  
nationality           0  
club                  0  
wage_eur              0  
preferred_foot        0  
international_reputation 0  
pace                 2036  
shooting              2036  
passing               2036  
dribbling             2036  
defending             2036  
gk_diving             16242  
gk_handling           16242  
gk_kicking            16242  
gk_speed              16242  
gk_reflexes           16242  
dtype: int64
```



```
[12]: # Replacing all the null values with 0
```

```
fifa.fillna(0, inplace=True)
```

```
fifa.isnull().sum()
```

```
[12]: short_name          0
```

```
age                    0
```

```
dob                   0
```

```
height_cm             0
```

```
weight_kg             0
```

```
nationality           0
```

```
club                  0
```

```
wage_eur              0
```

```
preferred_foot        0
```

```
international_reputation 0
```

```
pace                  0
```

```
shooting              0
```

```
passing               0
```

```
dribbling             0
```

```
defending             0
```

```
gk_diving             0
```

```
gk_handling           0
```

```
gk_kicking            0
```

```
gk_speed              0
```

```
gk_reflexes           0
```

```
dtype: int64
```

```
[13]: # Renameing the column
fifa.rename(columns={'wage_eur' : 'salary'}, inplace=True)
```

```
[14]: # Save the filtered DataFrame to a new CSV file
fifa.to_csv('filteredData.csv', index=False)
```

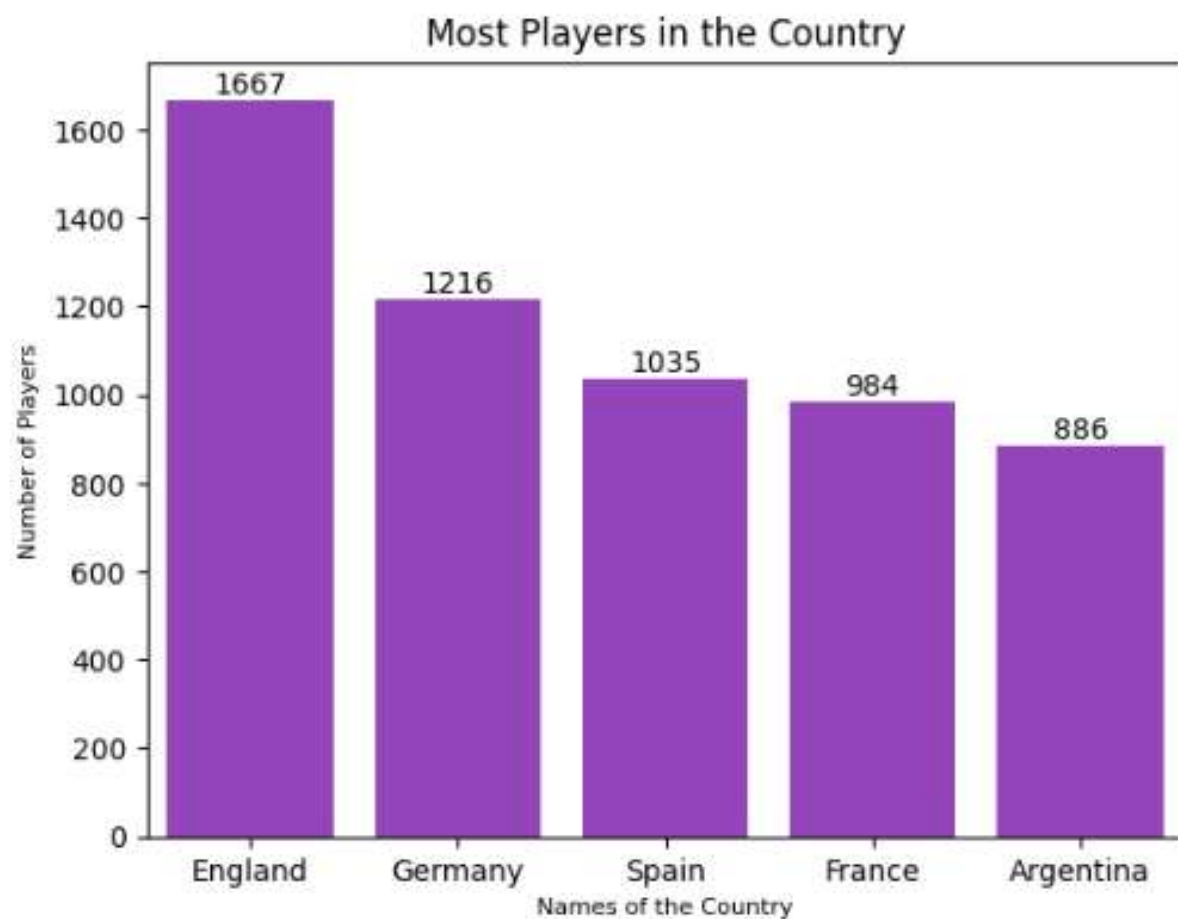
## DATA VISUALIZATION

### Which country has the most number of players (Top 5)?

```
[15]: # Top 5 countries and number of players in each country.
nationality_counts = fifa['nationality'].value_counts().head()
nationality_counts
```

```
[15]: nationality
England      1667
Germany      1216
Spain        1035
France        984
Argentina     886
Name: count, dtype: int64
```

```
[16]: # Bar-plot of top 5 countries with most number of players.  
ax = sns.barplot(x=nationality_counts.index, y=nationality_counts.values, color='DarkOrchid')  
for i in ax.containers:  
    ax.bar_label(i)  
plt.title('Most Players in the Country')  
plt.xlabel('Names of the Country',fontsize=8)  
plt.ylabel('Number of Players',fontsize=8)  
plt.show()
```



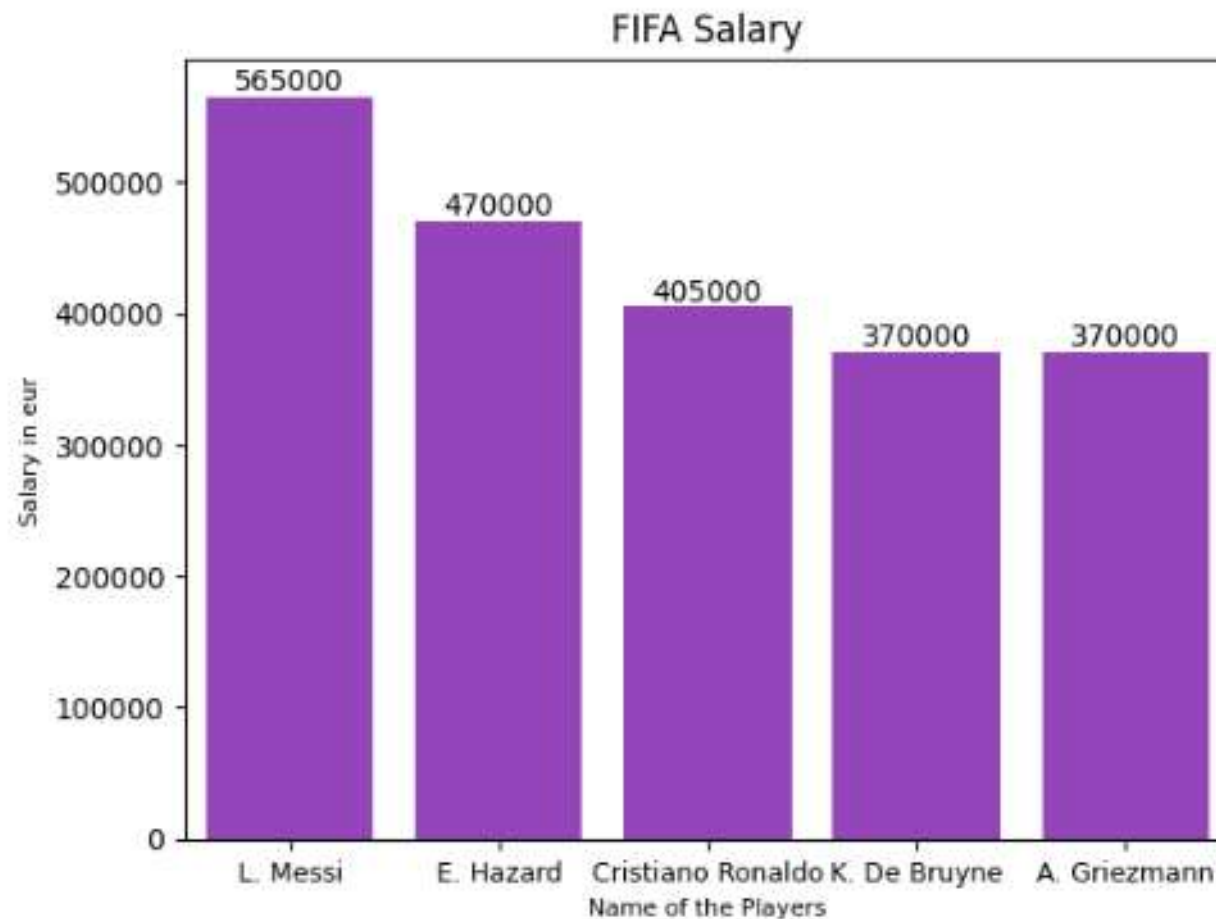
Which players has the highest salary ?

```
[17]: # Make a new dataset named player_salary of short_name and wage_eur columns.  
player_salary = fifa[['short_name','salary']]  
player_salary
```

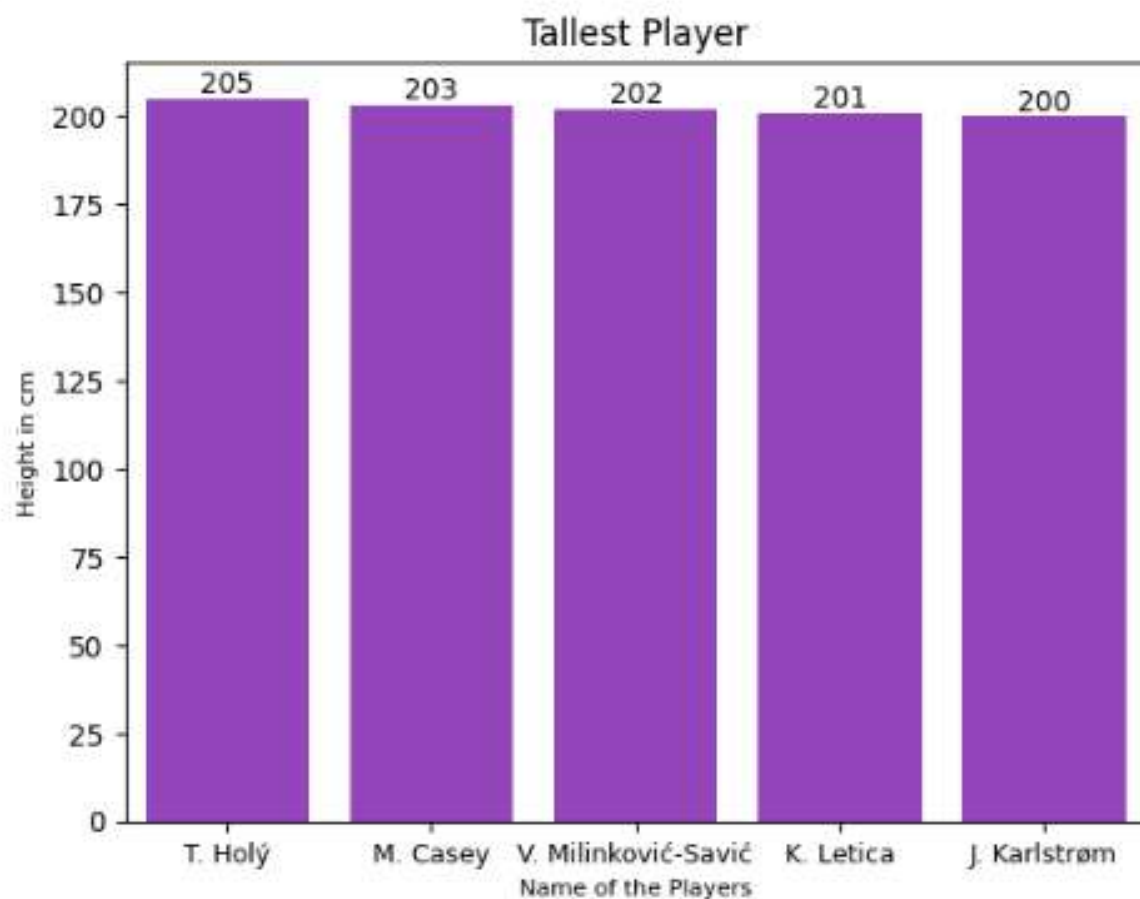
```
[17]:
```

	short_name	salary
0	L. Messi	565000
1	Cristiano Ronaldo	405000
2	Neymar Jr	290000
3	J. Oblak	125000
4	E. Hazard	470000
...	...	...
18273	Shao Shuai	2000
18274	Xiao Mingjie	2000
18275	Zhang Wei	1000
18276	Wang Haijian	1000
18277	Pan Ximing	2000

```
[19]: # Bar-plot of top 5 players with highest salary.  
ax = sns.barplot(x=sort_value['short_name'], y=sort_value['salary'], color='DarkOrchid')  
for i in ax.containers:  
    ax.bar_label(i)  
plt.title('FIFA Salary')  
plt.xlabel('Name of the Players',fontsize=8)  
plt.ylabel('Salary in eur',fontsize=8)  
plt.xticks(fontsize=9)  
plt.show()
```

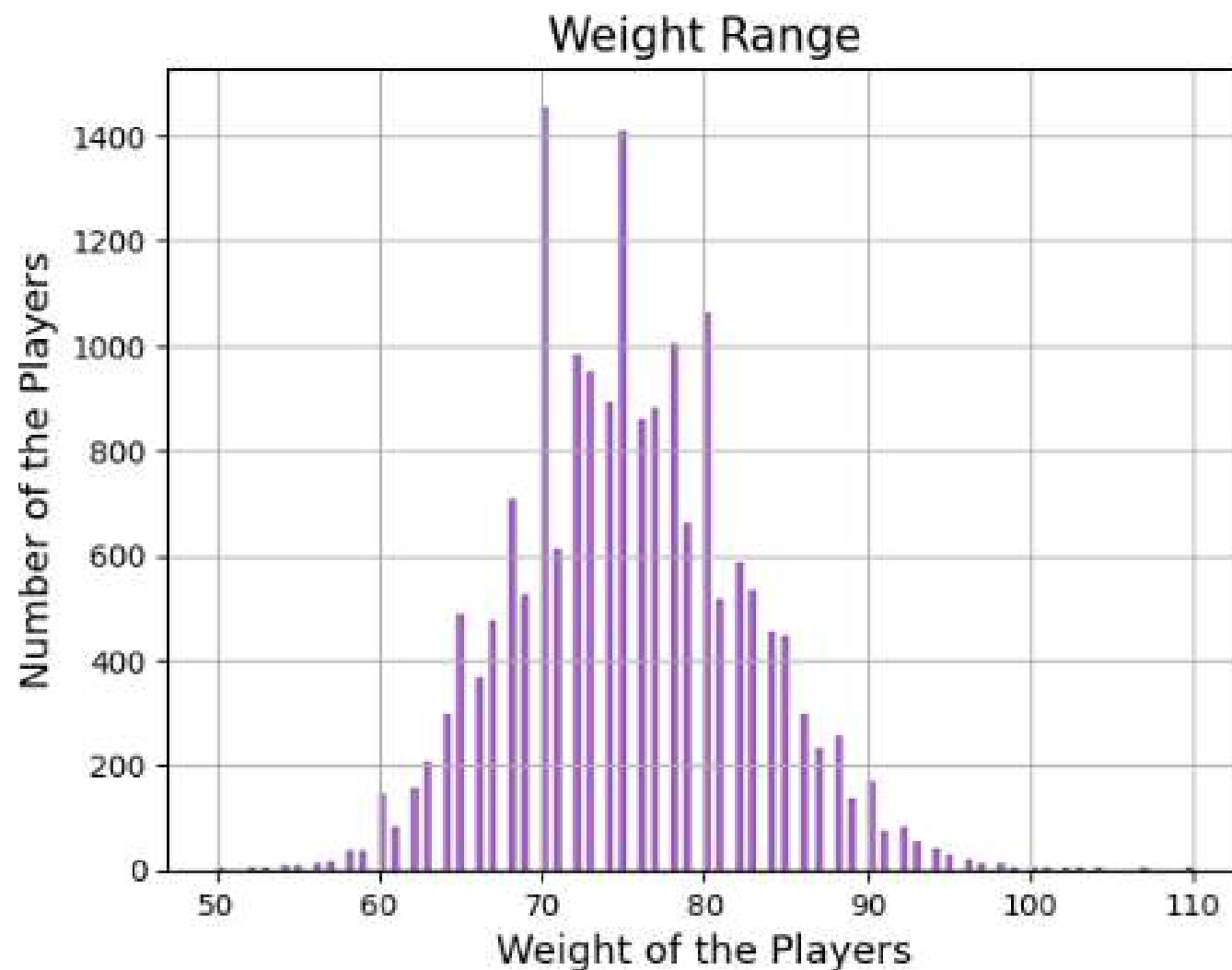


```
[22]: # Bar-plot of top 5 tallest player.  
ax = sns.barplot(x=sort_value['short_name'], y=sort_value['height_cm'], color='DarkOrchid')  
for i in ax.containers:  
    ax.bar_label(i)  
plt.title('Tallest Player')  
plt.xlabel('Name of the Players',fontsize=8)  
plt.ylabel('Height in cm',fontsize=8)  
plt.xticks(fontsize=9)  
plt.show()
```



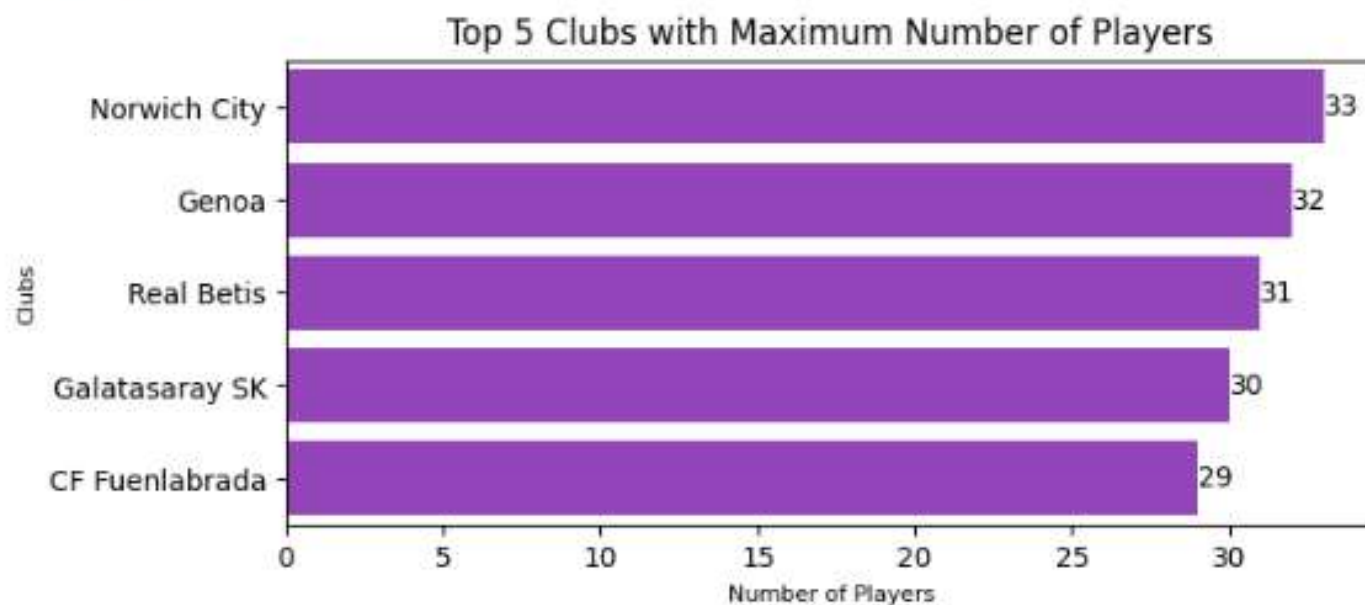
What is the range between weight and the players ?

```
[26]: # Bar-plot of top 5 heaviest player.  
ax = sns.histplot(player_weight['weight_kg'], bins=150, color='DarkOrchid')  
  
plt.title("Weight Range",fontsize=15)  
plt.ylabel('Number of the Players',fontsize=13)  
plt.xlabel('Weight of the Players',fontsize=13)  
plt.grid(True)  
plt.show()
```



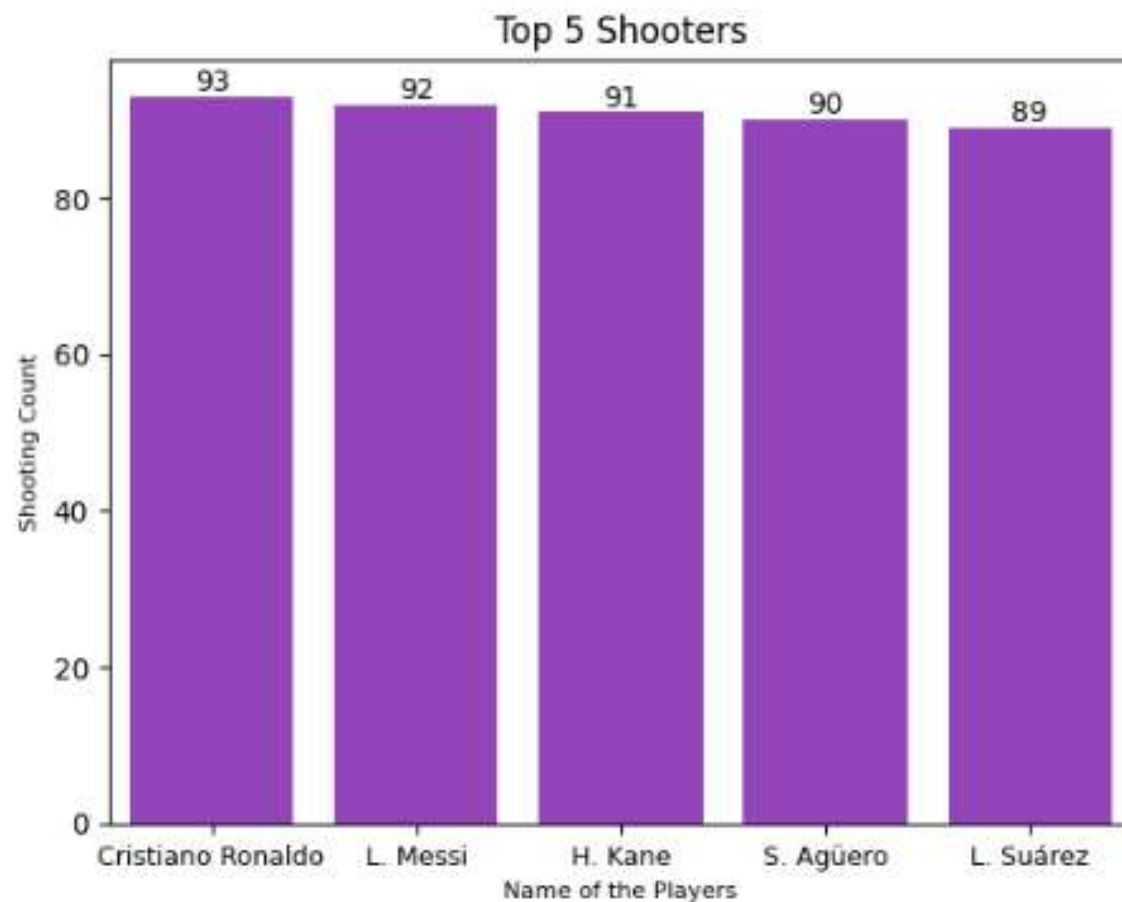
```
[29]: # Bar-plot of top 5 club who have the maximum number of players..
plt.figure(figsize=(7,3))

ax = sns.barplot(x=top_clubs.values, y=top_clubs.index, color='DarkOrchid')
for i in ax.containers:
    ax.bar_label(i)
plt.title('Top 5 Clubs with Maximum Number of Players')
plt.xlabel('Number of Players',fontsize=8)
plt.ylabel('Clubs',fontsize=8)
# plt.xticks(fontsize=9)
plt.show()
```





```
[32]: # Bar-plot of top 5 Shooting player.  
ax = sns.barplot(x=sort_value['short_name'], y=sort_value['shooting'], color='DarkOrchid')  
for i in ax.containers:  
    ax.bar_label(i)  
plt.title('Top 5 Shooters')  
plt.xlabel('Name of the Players',fontsize=8)  
plt.ylabel('Shooting Count',fontsize=8)  
plt.xticks(fontsize=9)  
plt.show()
```



## Insights

- 💡 England has the most players 1667.
- 💡 L.Messi has the highest salary €565,000
- 💡 T. Holý is the tallest player 205cm
- 💡 A. Akinfenwa is the heaviest player with 110Kg
- 💡 Most of the players have weight between 70-80 kg.
- 💡 Norwich City has the most number of players with 33 players.
- 💡 Cristiano Ronaldo is the top shooter with 93 shots.
- 💡 G. Chiellini is the best defender 90 counts.
- 💡 L. Messi is the best dribbler 96 counts.
- 💡 De Gea is the best goalkeeper with 92 reflexing counts.