

به نام خدا

دکتر محمدزاده

آزمایشگاه یادگیری و بینایی ماشین

تمرین سری اول

تمرین اول:

داده

داده‌ای که در این بخش استفاده می‌شود، اطلاعات عملکرد پایان فصل تیم‌های فوتبال لیگ اسپانیا است از سال ۱۹۷۰ تا ۲۰۱۶. داده‌ها را می‌توانید از این [لینک](#) دانلود نمایید. در هر سطر، اطلاعاتی از قبیل، تعداد برد-باخت-تساوی، مجموع گل‌های زده-خورده، مجموع امتیازها پایان فصل و ... به ازای هر تیم در هر فصل، آورده شده است.

در ادامه سعی می‌کنیم با توجه به اطلاعات در دسترس قرار گرفته، بررسی مختصری در مورد داده انجام بدهیم.

ابتدا داده‌های به فرمت csv را با استفاده از کتابخانه pandas به داخل یک دیتافریم ریخته و با استفاده از توابع head یا tail می‌توانید چند سطر اول و آخر را مشاهده نمایید.

سپس با استفاده از تابع dtype می‌توانید تایپ مربوط به هر کدام از ستون‌ها را مشاهده نمایید. خروجی نشان دهنده نوع ذخیره سازی هر ستون توسط pandas می‌باشد.

اطلاعات اولیه آماری ستون‌ها

می‌توانید با استفاده از تابع describe شمای کلی‌ای از توصیف آماری داده‌ها بدست آورید.

	home_win	away_win	home_loss	away_loss	matches_won	matches_lost	matches_drawn	total_matches	points	home_goals	away_goals	goals_scored	g
count	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	908.000000	
mean	9.582599	4.118943	4.118943	9.582599	13.701542	13.701542	9.577093	36.980176	50.681718	29.508811	18.267621	47.776432	
std	3.107555	2.776725	2.531206	3.092974	4.966050	4.589442	2.986852	2.522150	14.104626	9.757192	7.569011	15.593429	
min	1.000000	0.000000	0.000000	1.000000	2.000000	1.000000	1.000000	30.000000	13.000000	11.000000	4.000000	15.000000	
25%	7.000000	2.000000	2.000000	8.000000	10.000000	11.000000	7.000000	34.000000	42.000000	23.000000	13.000000	37.000000	
50%	9.000000	4.000000	4.000000	10.000000	13.000000	14.000000	9.000000	38.000000	48.000000	28.000000	17.000000	45.000000	
75%	12.000000	6.000000	6.000000	12.000000	16.000000	17.000000	12.000000	38.000000	59.000000	34.000000	22.000000	54.000000	
max	19.000000	16.000000	15.000000	18.000000	32.000000	29.000000	18.000000	44.000000	100.000000	78.000000	58.000000	121.000000	

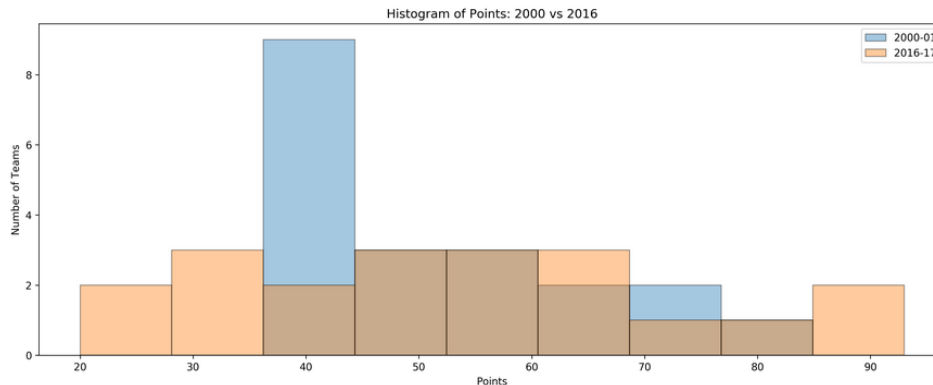
برای داده‌های از نوع عدد ستون‌ها به صورت زیر می‌باشند.

- count: اعداد ردیف‌هایی که مقداری در آن وجود دارد (گاهی ممکن است برخی از مقدارهای یک ستون، وارد نشده باشند!)
- mean: میانگین مقدارهای آن ستون
- std: انحراف معیار مقدارهای آن ستون
- min و max: به ترتیب «کمترین مقدار» و «بیشترین مقدار» در آن ستون

• ۰.۲۵ و ۰.۵۰ و ۰.۷۵: به ترتیب چارک های مقدار های آن ستون

اما برای بررسی بهتر مقدار های ستون ها میخواهیم ببینیم آیا نمودار توزیع تیم ها در پایان لیگ ۲۰۰۰ و ۲۰۱۶ با هم تفاوت دارند؟
تمرین: برای فهم این موضوع ابتدا نمودار هیستوگرام امتیاز تیم ها در این دو فصل را رسم کنید.

نمودار شما باید شکلی همانند شکل زیر داشته باشد.



همانگونه که ملاحظه میکنید تفاوت قابل توجهی بین این دو هیستوگرام وجود دارد. در سال ۲۰۰۰ امتیاز ها به همدیگر نزدیک تر بودند. در سال ۲۰۱۶، امتیازها متنوع تر و با تفاوت بیش تری هستند.

یک راه برای بررسی ارتباط تنوع امتیازات (به شکل کیفی و تصویری)، رسم کردن هیستوگرام تمام فصل ها از ۲۰۰۰ تا ۲۰۱۶ است. سپس باید به تصویر(ها) نگاه کرد و به دنبال وجود یک روند تغییر در هیستوگرام بود. سپس سعی بر کمی (تبدیل به عدد کردن) این مشاهده ها کرد. یعنی اگر به این نتیجه رسیدیم که تفاوتی در هیستوگرام ها وجود دارد، معیاری تعریف کرد که این تفاوت را به عدد تبدیل کند. یا حتی اگر تفاوتی مشاهده نشد، معیاری تعریف کرد که عدم تفاوت را نشان دهد.

تمرین: دو تابع `min_point` و `max_point` را پیاده سازی کنید به طوری که با گرفتن `datafram` لالیگا به عنوان آرگومان اول و نام فصل به عنوان آرگومان دوم دیکشنری حاوی نام و امتیاز تیم صدر و انتهای جدول را برگرداند. به عنوان نمونه

```
>>> df_laliga = pd.read_csv('LaLiga_dataset.csv')
>>> min_points(df_laliga, '1997-98')
{'club': 'Sporting de Gijon', 'points': 13}
>>> max_points(df_laliga, '1997-98')
{'club': 'Barcelona', 'points': 74}
```

قهرمان فصل

در تمرین قسمت قبل دیدید که ابتدا با مرتب کردن تیم ها بر اساس امتیاز و سپس گروه بندی آنها بر اساس فصل می توان جدول هر فصل (و طبیعتاً قهرمان آن را) پیدا کرد. حال بیایید ببینیم قهرمان لالیگا چقدر تنوع دارد.

میخواهیم بررسی کنیم که قهرمان لالیگا از سال ۱۹۷۰ تا به حال، چه تیم‌هایی بوده‌اند؟ هر تیم چند دفعه قهرمان «لالیگا» شده است؟ یا اینکه چند تیم تا به حال توانسته‌اند قهرمان لیگ اسپانیا شوند؟ بنظر میرسد برای پاسخ دادن به این سوال نمودار میله‌ای مناسب باشد.

تمرین

در این بخش از شما می‌خواهیم:

- (۱) پیدا کردن جدول هر فصل
- (۲) پیدا کردن قهرمان هر فصل
- (۳) پیدا کردن تعداد قهرمانی‌های هر تیم و اینکه هر تیم چندبار قهرمان شده
- (۴) تعداد تیم‌های موجود در هر فصل

تذکر:

- (۱) فایل `club_count.csv` دارای دو ستون؛ یکی فصل‌ها و دیگری تعداد تیم‌ها باشد.
- (۲) فایل `champions.csv` باید دارای دو ستون؛ یکی فصل‌ها و دیگری قهرمان‌ها باشد.
- (۳) فایل `championship_count.csv` باید دارای دو ستون؛ یکی باشگاه و دیگری تعداد قهرمانی‌ها باشد.
- (۴) به هیچ وجه از دستور `for` برای کار با `pandas` استفاده نشود.