به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

**درس استنتاج آماری**

**پروژه (فاز یک)**

**محمد بزم**

**810196126**

اردیبهشت ماه 1397

**فهرست**

|  |  |
| --- | --- |
| عنوان | شماره صفحه |
| چکیده | ۳ |
| تمرین 1 | 4 |
| تمرین 3 | 6 |
| تمرین 4 (تمرین عملی) | 8 |
| پیوست 1 : روند اجرای برنامه | 11 |

مقدمه:

توسط این دادگان تصمیم‌سازی فنی (انتخاب تکنیک مناسب برای بازی) با استفاده از مدل‌های اطلاعاتی‌ ساخته شده توسط دانش یادگیری ماشین انجام می‌شود. برنامه‌ریزی برای نقل و انتقال بازیکنان شامل تصمیم‌گیری در مورد جذب یا عدم جذب آنان از دیگر تصمیماتی است که بر اساس نتایج آنالیز داده گرفته می‌شود. از موارد دیگر می‌توان به رتبه بندی بازیکنان بر اساس معیارهای مختلف با استفاده از آنالیز دادگان مربوط به آنان اشاره کرد. برای مثال برای انتخاب بازیکنان موثر در خط حمله با استفاده از روش‌های ساده تنها با پارامترهایی مانند تعداد گل زده و یا تعداد پاس گل داده شده انجام می‌شود اما با روش‌های مبتنی بر آنالیز دادگان پارامترهای تاثیرگذار بیشتری را می‌توان پیدا کرد که شاید کشف آنها با روش‌های عادی امکانپذیر نباشد و تحلیل دقیق‌تری را می‌توان ارائه کرد.

دادگان (dataset) فیفا دربردارنده‌ی دانسته‌هایی در مورد بازیکنان مختلف فوتبال است. این مجموعه شامل 185 ویژگی مختلف است که از میان آنها به موارد زیر اشاره کرد:

1. دانسته‌های مربوط به بازیکنان شامل: سن، باشگاه، لیگ، ملیت، درآمد، ویژگی‌های فیزیکی و ...

2. ویژگی‌های مربوط به بازی آنها شامل: دریبل و ...

3. ویژگی‌های منحصر به فرد نظیر یک حرکت شاخص یا شهرت.

ویژگی‌های مورد اشاره در این بررسی را می‌توان در چند بخش گنجاند. یک بخش دانسته‌هایی را شامل می‌شوند که خود بازیکن را توصیف می‌کنند (قد، وزن و . . .)، بخش دیگر هویت ورزشی آنان است (شهرت، ویژگی‌های شاخص بازی و . . .)، بخشی عملکرد آنان رو توصیف می‌کند (دریبل، کنترل توپ و . . .).

# سوال 0

بررسی نسبت مشاهدات null به کل مشاهدات در بیشتر ویژگی‌ها صفر بود. اما در 28 تا از ویژگی‌ها این طور نمی‌باشند نام آنها به شرح زیر است. بررسی به عمل آمده نشان می‌دهد برای تمام این ویژگی‌ها نسبت مورد نظر برابر با 1123/0 می‌باشد.

نسبت مورد نظر توسط کد زیر بدست می‌آید:



جدول 1: ویژگی‌هایی که نسبت تعداد مشاهدات تهی به غیر تهی برای آنها غیر صفر شده است.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ویژگی‌های با نسبت غیر صفر تهی با کل مشاهدات** | | | | | | | |
| Lf | 22 | Cm | 15 | Rm | 8 | Rs | 1 |
| Lam | 23 | Lm | 16 | Rb | 9 | Rw | 2 |
| Lcb | 24 | Cdm | 17 | Rwb | 10 | Rf | 3 |
| Ldm | 25 | Lwb | 18 | St | 11 | Ram | 4 |
| Lcm | 26 | Lb | 19 | Cam | 12 | Rcm | 5 |
| Gk | 27 | Cb | 20 | Cf | 13 | Rcb | 6 |
| ls | 28 | ls | 21 | lw | 14 | rdm | 7 |

این مشاهدات بدون مقدار را می‌توان به عنوان مثالی از non responding در نظر گرفت. چون تعداد این مشاهدات نسبت به تعداد کل مشاهدات زیاد نیست، می‌توان آنها را حذف کرد بدون اینکه به نتایج تحلیل خدشه‌ای وارد شود.

# سوال 1

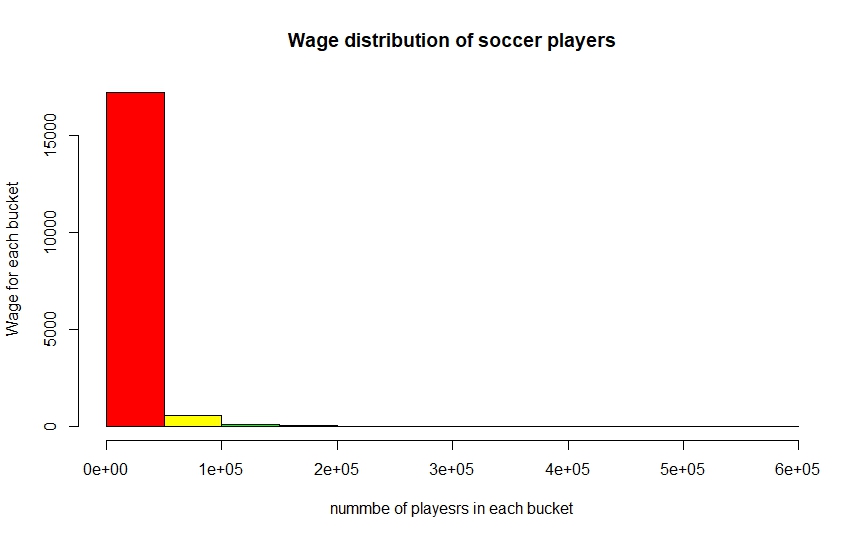
برای این کار، با استفاده از دستور زیر، درآمد بر حسب یورو (eur\_wage) انتخاب شد.



## بخش 1:

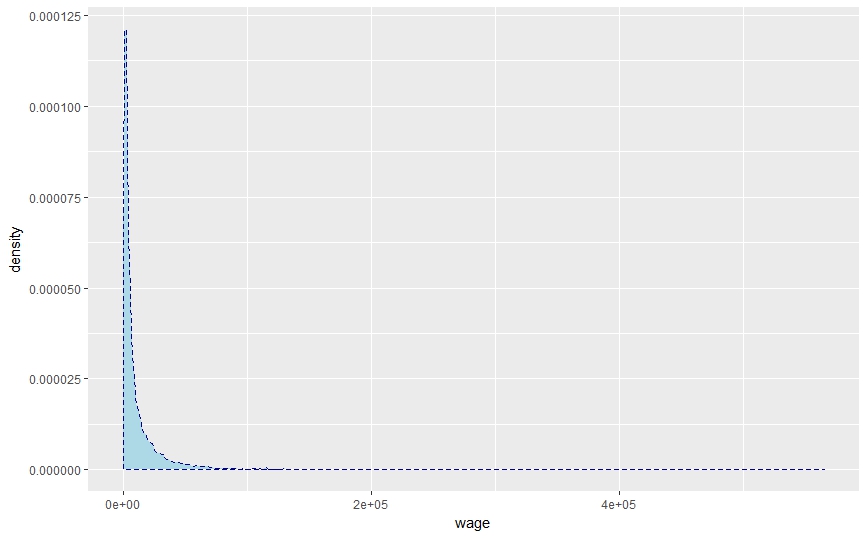
در این بخش با استفاده از دستورات زیر نمودار خواسته رسم شده است. همچنین تعداد bin ها برابر با 10 در نظر گرفته شده است. به این دلیل که با این تعداد می‌توان بیشترین تعداد ستون را مشاهده کرد یعنی با بیشتر کردن تعداد bin ها ستون جدیدی اضافه نمی‌شود چون داده‌ای در آن قرار نمی‌گیرد.





شکل 1 : هیستوگرام درآمدر بازیکنان

## بخش 2:



شکل 2 : چگالی داده‌های مربوط به متغیر wage

رسم این نمودار توسط کد زیر انجام شده است:



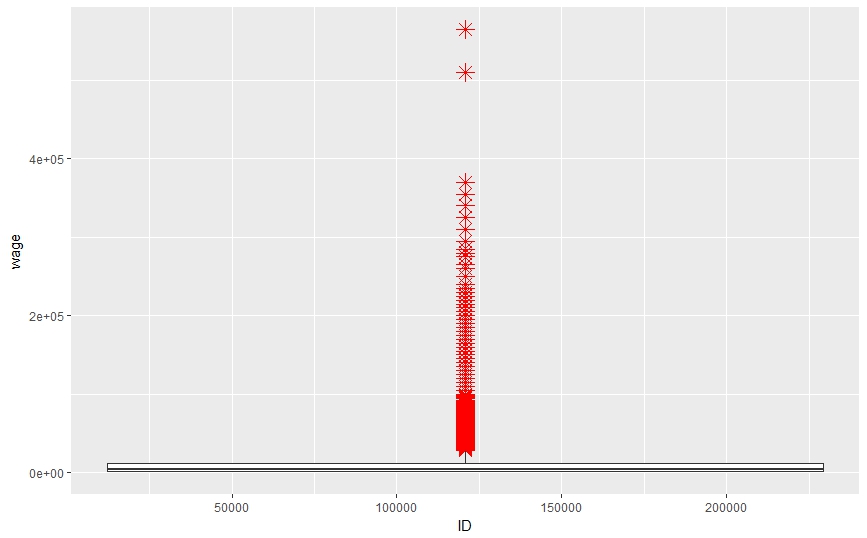
## بخش 3:

همانطور که از نمودارهای چگالی و هیستوگرام مشخص است، این نمودار به شدت به سمت راست چوله است و نیز تنها یک مُد دارد.

## بخش 4:

|  |  |  |  |
| --- | --- | --- | --- |
| **Skewness** | **STD** | **Variance** | **Mean** |
| 0.3255 | 23050.66 | 531332976 | 11503.83 |

## بخش 5:



شکل 3 : نمودار Boxplot برای متغیر wage .

جزئیات این نمودار به شرح زیر است:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Max | 3rd Quantile | Median | 1st Quantile | Min |
| 565000 | 12000 | 4000 | 2000 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Upper Outer fence | Lower Outer fence | Upper Inner fence | Lower inner fence | IQR(Q3-Q1) |
| 32000 | -28000 | 27000 | -13000 | 10000 |

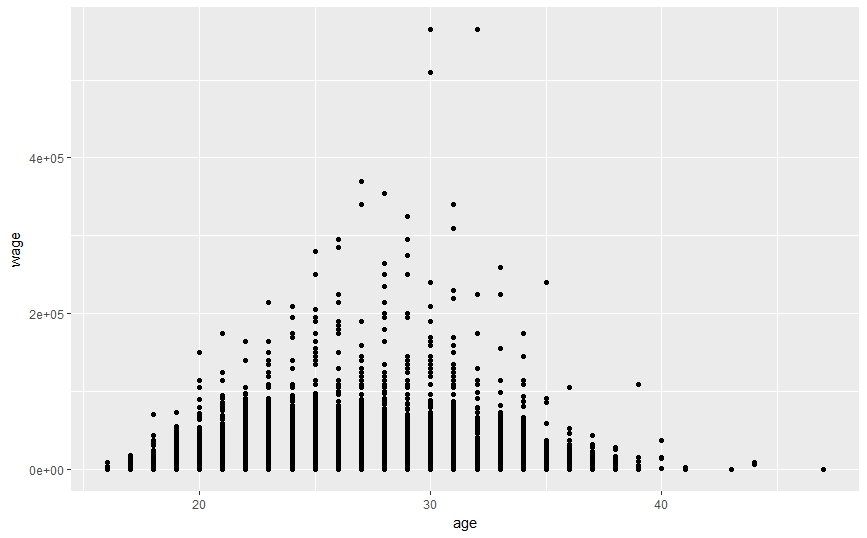
## بخش 6:

با توجه به نمودار جعبه‌ای، هیچ outlier ای در قسمت پایین وجود ندارد. همچنین عنصر کمینه برابر با صفر است و Lower Inner Fence در قسمت اعداد منفی قرار می‌گیرد که این هم موید عدم وجود عنصر outlier در قسمت پایین می‌باشد. اما در قسمت بالا outlier هایی وجود دارند که مقدار و تعداد آنها توسط کد زیر قابل محاسبه است.



# سوال 2:

در این بخش متغیر‌های درآمد (wage) ، سن (Age) و قد (height) در نظر گرفته می‌شوند. هدف از انتخاب این دو متغیر، بررسی تاثیر عوامل افزایش سن و قد بر میزان درآمد بازیکنان فوتبال است.

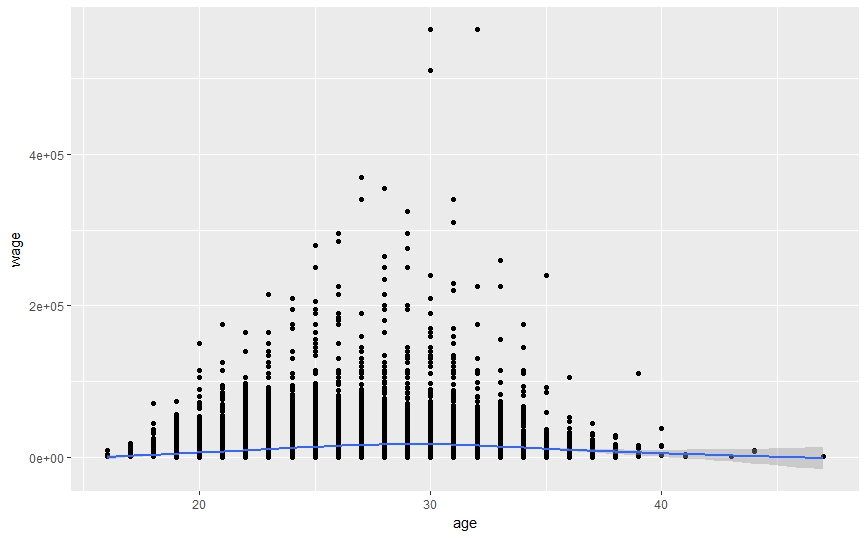


شکل 4 : scatterplot برای دو متغیر age و wage

همانطور که از نمودار مشخص است رابطه‌ی این دو متغیر غیر خطی است.

معیار correlation coefficient برای این دو متغیر برابر با 0.15 می‌باشد. این عدد به معنی وجود رابطه‌ی خطی ضعیف اما مثبت بین دو متغیر می‌باشد.

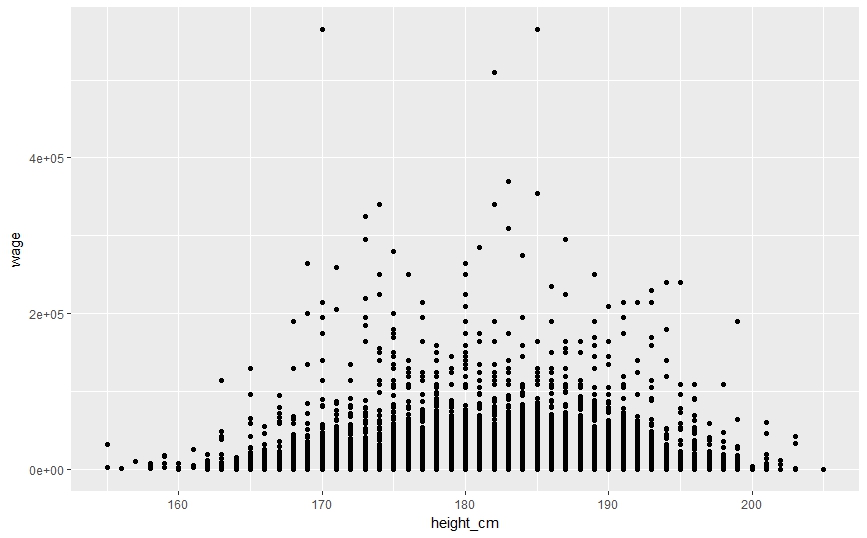
## بخش 4:



شکل 6 : تقریب یک خط بر روی نمودار scatterplot

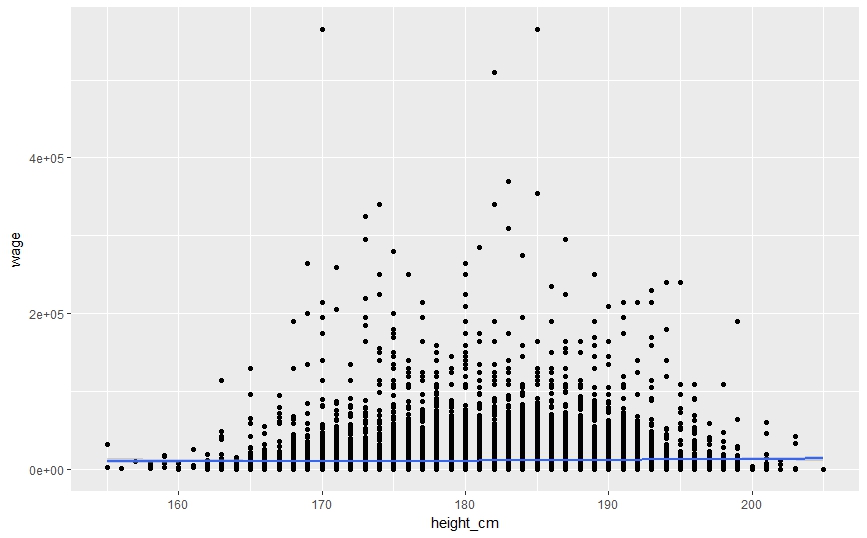
خط آبی رنگ زیر نمودار، همان خطی اس که بر روی نمودار تقریب زده شده است. این خط مانند توزیع نمونه‌ها تا حدود 28 شیب مثبت و پس از آن شیب منفی دارد. شیب خط همواره مقداری کوچک دارد که این مشاهده خود موید عدد بدست آمده از Correlation coefficient می‌باشد.

## بخش 5:



شکل 7 : نمودار scatterplot برای دو متغیر درآمد و قد

معیار Correlation coefficient برای دو متغیر درآمد و قد برابر با 0.02 می‌باشد. این عدد به معنی وجود رابطه‌ی خطی ضعیف اما مثبت بین دو متغیر می‌باشد.



شکل 9 : نمودار scatterplot به همراه خط تخمین زده شده برای دو متغیر درآمد و قد.