هدف این پروژه استفاده از مباحث و مفاهیم مطرح شده در این درس جهت تحلیل اطلاعات در داده‌هایی واقعی است . چی داده‌ای؟ این با شما خواهید بود!

سایت kaggle.com بستر بسیار مناسبی را برای دسترسی محقیقن و مهندسین به دادههای واقعی فراهم آورده است. ایده اصلی این سایت برگزاری رقابتهای تحلیل داده است، شرکتهای مختلف با همکاری این سایت داده‌های خود را در اختیار اعضای سایت قرار میدهند .هر شرکت هدف مورد نظر خود را بیان می‌کند و تحلیلی خاص را طلب میکند و افراد علاقمند تلاش میکنند بهترین تحلیل را انجام دهند. برنده چنین رقابتی معمولا پاداش دریافت می‌کند بعضی از این پاداشها بسیار قابل توجه هستند!

داده مورد علاقه خود را انتخاب کنید، برای مثال می‌توانید داده‌های زیر، که به نظر تیمی تدریسیاری جذاب بوده‌اند، انتخاب کنید:

https://www.kaggle.com/hugomathien/soccer(European Soccer Database)

* Imdb data set(include many movie related measurements like its rating on the imdb site, the cast of the movie, how much it grossed and …)

https://www.kaggle.com/dgawlik/nyse(New York Stock Exchange)

*داده‌های سرشماری ایران (شامل پارامترهایی مثل وضعیت شغلی، دسترسی به اینترنت، وضعیت محل سکونت و ...)

*این داده ها در سایت درس بارگذاری خواهند شد.

۰. در ابتدا بیان کنید چرا دادهای که انتخاب کرده‌اید برایتان جذاب است؟تفسیرتان از داده‌ی انتخاباتیان چیست؟

۱. یک متغیر عددی را انتخاب کرده و هیستوگرام مربوط به آن را ترسیم کرده، madality و Skewness آن را تحلیل کنید. برای همین متغیر boxplot را ترسیم کرده و تعداد outlier ها را مشخص کنید.

۲. دو متغیر عددی دلخواه انتخاب کنید و رابطه‌ی بین آنها را در قالب یک scatterplot نمایش دهید سپس این نمودار را تحلیل کنید. کورولویشن و کورایانس این دو متغیر را بدست آورید.

۳. یک متغیر رسته‌ای (Categorical) را به دلخواه انتخاب کرده و جدول فرکانسی و نمودار میله‌ای آن را ترسیم کنید.

۴. دو متغیر رشتهای انتخاب کنید و برای این دو Contingency Table، Segmented Bar Plot و Mosaic Plot را نمایش دهید.

۵. یه متغیر عددی انتخاب کنید و برای میانگین آن بازه اطمینان ۹۹ درصد را محاسبه کنید. سپس آن را تحلیل کنید.

۶. برای میانگین یک متغیر عددی، آزمون فرضی را مطرح کنید و با محاسبه p-value فرض خود را تایید و یا رد کنید.

۷. یک متغیر پاسخ (response) عددی و تعدادی متغیر توضیحی عددی و رسته ای انتخاب کرده و معادله خط رگرسیون را محاسبه کنید. شیب همه متغیرهای توضیحی و عرض از مبدا را تفسیر کنید. برای حداقل یکی ازمتغیرهای توضیحی آزمون فرضی طراحی و اجرا کنید و نتیجه بگیرید آیا این متغیر توضیحی برای پیشبینی متغیر پاسخ مناسب است یا خیر.

گزارش نهایی شما باید شامل پاسخ دقیق و کامل به سوالات فوق‌الذکر و کلیه کدهای R مورد استفاده باشد. با توجه به میزان خلاقیت در پاسخگویی به سوالات و ترسیم نمودارها، صحت و دقت و جامعیت تحلیلها امکان نمره اضافی برای شما فراهم است.