تکلیف 2 مقدمه ای بر علم داده

AT(ها): شهرزاد جاویدی، ملیکا صادقی
مدرسان: دکتر بهرک، دکتر یعقوب زاده

مهلت: جمعه 17فروردین
ساعت 23:59

معرفی

این تکلیف شامل وظایف خاصی برای بررسی سوالات باز است. پرسش‌های باز از شما می‌خواهند که خلاقانه و انتقادی در مورد اینکه چگونه طرح‌هایی که ایجاد کرده‌اید بینشی در مورد داده‌ها ارائه می‌دهند، فکر کنید.

وظیفه 1

TA:ملیکا صادقی مجموعه داده ارائه شده (task1.csv)حاوی اطلاعاتی درباره مسافران کشتی غرق شده "RMS Lusitania"است.
در این کار با کار با pandas ،numpyو matplotlibآشنا می شوید. این توضیح مختصری از ستون‌های مجموعه داده داده شده است:

Surved:•نشان می دهد که آیا یک مسافر زنده مانده است (1)یا نه (0). • pclass: (1)کلاس بلیط = 2کلاس اول، = 2کلاس دوم، = 3کلاس سوم). •جنسیت: جنسیت مسافر (مرد یا زن). •سن: سن مسافر بر حسب سال. sibsp:•تعداد خواهر و برادر یا همسری که مسافر در کشتی داشته است. parch:•تعداد والدین یا فرزندانی که مسافر در کشتی داشته است. •کرایه: کرایه ای که مسافر برای بلیط پرداخت کرده است. embarked:•بندری که مسافر در آن سوار شد (C = Cherbourg، Q = Queenstown، S = Southampton). •کلاس: کلاس بلیطی که مسافر داشته است (اول، دوم یا سوم). who:•مسافران را به‌عنوان «مرد»، «زن» یا «کودک» دسته‌بندی می‌کند که احتمالاً از سن و ارتباط جنسی

adult_male:•یک بولی که نشان می دهد مسافر مرد بالغ است یا خیر. •عرشه: عرشه ای که کابین مسافر روی آن قرار داشت که با حروف مشخص شده بود. embark_town:•شهری که مسافر از آنجا سوار شده است، مطابق با کدهای 'embarked'.

•زنده: نشان می دهد که آیا مسافر زنده مانده است ("بله") یا نه ("نه")

•تنهایی: یک بولین که نشان می‌دهد مسافر تنها سفر می‌کرده است (بدون خواهر و برادر، همسر، والدین یا فرزند).

سوالات

.1ابتدا فایل را با استفاده از کتابخانه pandasبخوانید و آن را در یک dataframe pandasذخیره کنید. سپس با استفاده از روش‌های 'اطلاعات'، 'سر'، 'دم'، 'توضیح' از کتابخانه پانداها، ساختار کلی داده‌ها را بررسی کنید و توضیح دهید که هر یک از خروجی‌ها چه اطلاعاتی را نشان می‌دهند.

.2نوع هر ستون داده را نشان دهید. برخی از ستون‌ها از نوع طبقه‌بندی و برخی از نوع عددی از اطلاعات کتابخانه پانداها هستند. برای پردازش ستون های غیر عددی، یکی از روش های ممکن برچسب گذاری است. به گونه ای که هر یک از دسته ها با یک عدد جایگزین می شود. به عنوان مثال، در این مجموعه داده، یک ستون دسته بندی به نام sexوجود دارد که شامل مقادیر Maleو Femaleاست. مقادیر این ستون را تغییر دهید، به طوری که هر یک از این مدل ها به یکی از اعداد در محدوده [0، 1]نگاشت شوند.

.3یک نقشه حرارتی از ماتریس همبستگی برای ویژگی های عددی در مجموعه داده ایجاد کنید.

.4ستون هایی که روی مورب اصلی نیستند و همبستگی 1دارند به این معنی است که یک ستون اضافی است و قابل حذف است. بنابراین، بر این اساس، ستون های اضافی که این شرایط را دارند حذف کنید.

.5چند مسافر از فاجعه جان سالم به در بردند ؟(1 ==)

.6همه مسافران زن بالای 30سال را پیدا کنید. چند نفر هستند؟

.7مسافرانی را که از شربورگ ("C")سوار شده‌اند و کرایه‌ای بیشتر از 100دلار

.8ستون هایی را با مقادیر گمشده شناسایی کنید. یک استراتژی برای مدیریت این موارد پیشنهاد و اعمال کنید ارزش از دست رفته.

.9میانگین سنی مسافران کشتی چقدر است؟ چه تفاوتی بین مردان دارد و زنان؟

.10آیا بین کرایه پرداختی و نرخ بقا همبستگی وجود دارد؟ یک آمار ارائه دهید خلاصه.

.11از Matplotlibبرای ترسیم نسبت مسافرانی که بر اساس کلاس زنده مانده اند استفاده کنید.

.12توزیع سنی مسافران را ترسیم کنید و بین کسانی که زنده مانده اند تمایز قائل شوید و کسانی که این کار را نکردند.

.13یک نمودار پراکنده ایجاد کنید که رابطه بین سن و کرایه پرداخت شده را با کد رنگی نشان می دهد. با بقا

.14یک جدول محوری برای نشان دادن میانگین کرایه و نرخ بقا برای هر کلاس و جنس ایجاد کنید

ترکیبی

.15یک نمودار میله ای گروه بندی شده با استفاده از Matplotlibترسیم کنید تا میانگین کرایه پرداخت شده توسط مسافران، گروه بندی شده بر اساس کلاس و وضعیت بقای آنها را نشان دهد.

وظیفه 2

تا: شهرزاد جاویدی به این مجموعه داده (task2.csv)خوش آمدید، که بر حقوق دانشمندان داده در مناطق مختلف از سال 2020تا 2024تمرکز دارد. ما با هم به بررسی بینش آن خواهیم پرداخت. بیا شروع کنیم!

سؤالات •در ابتدا، بهتر است یک پیش تحلیل انجام دهید تا در صورت وجود، داده های تکراری و NA(مفقود شده) را حذف کنید. سپس، از آنجایی که حقوق ها ذاتاً به واحد پول هر کشور مربوطه گره خورده است، برای مقایسه های معنادار باید آنها را به یک ارز واحد استاندارد کنیم. بیایید با شناسایی ارزهای موجود در مجموعه داده شروع کنیم. با توجه به ارزهای زیاد، بیایید بسامدهای آنها را بررسی کنیم و داده های مرتبط با ارزهایی که کمتر از ده بار نمایش داده شده اند را حذف کنیم.

•در این مرحله، این ارزها را به USDتبدیل می کنیم. شما می توانید این کار را به دو صورت انجام دهید: جستجوی دستی نرخ ارز از طریق منابع آنلاین مانند Googleیا استفاده از بسته های نرم افزاری مانند Forex-Pythonبرای تبدیل ساده. •اکنون، با بهره‌گیری از بینش‌های جمع‌آوری‌شده از مجموعه داده‌های شما، بیایید از تکنیک‌های مختلف تجزیه و تحلیل داده‌های اکتشافی (EDA)برای استخراج بینش‌های ارزشمند استفاده کنیم. برای مثال، می‌توانیم 10عنوان شغلی پرطرفدار یا 10بالاترین حقوق را شناسایی کنیم.

به امید موفقیت های زیادتان!

•توجه به این نکته مهم است که تجزیه و تحلیل و تفسیر کامل یافته‌ها و تجسم‌های شما برای ارزیابی جامع در این بخش بسیار مهم است. •نکته: بر اساس آنچه تاکنون آموخته اید، باید از تکنیک های تجسم متفاوت استفاده کنید. همچنین، باید توزیع متغیرها را بررسی کنید.

این شما هستید که تصمیم می گیرید کدام نمودار برای کدام متغیر(ها) مناسب است. اگر اطلاعات کافی از این مجموعه داده به دست آورده باشید، وظیفه شما "موفق" تلقی می شود.

یادداشت

•کار خود را به صورت فایل فشرده در این قالب در وب سایت آپلود کنید: .DS_CA2_[Std number].zip

اگر پروژه به صورت گروهی انجام می شود، تمام شماره دانشجویی اعضای گروه را در نام ذکر کنید.

•اگر پروژه به صورت گروهی انجام شود، فقط یک عضو باید اثر را آپلود کند. •ما کد شما را در حین تحویل پروژه اجرا خواهیم کرد، بنابراین مطمئن شوید که نتایج شما درست است

قابل تکرار