**شکافتن داده‌های دیوار: پروژه‌ای عملی در تحلیل بازار املاک**

**مقدمه**

در این پروژه جذاب، گامی در مسیر تحلیل داده‌های دنیای واقعی برمی‌داریم. با بهره‌گیری از مجموعه داده ارزشمندی که توسط شرکت محترم دیوار در اختیار ما قرار گرفته است، به کاوش در اعماق اطلاعات نهفته در تبلیغات املاک خواهیم پرداخت. این تجربه عملی، فرصتی بی‌نظیر برای به‌کارگیری دانش و مهارت‌های تحلیل داده در یک سناریوی واقعی را فراهم می‌آورد.

هدف اصلی این پروژه، فراتر از صرفاً انجام تحلیل‌های روتین، بستری برای توانمندسازی علاقه‌مندان به حوزه تحلیل داده و علم داده است. از یک سو، تلاش خواهیم کرد تا مجموعه‌ای از فعالیت‌های کلیدی که یک تحلیلگر داده باید بر آن‌ها مسلط باشد را به صورت عملی تجربه کنیم. این شامل مراحلی چون درک داده‌ها، پاکسازی و آماده‌سازی، انجام تحلیل‌های اکتشافی، مدل‌سازی (در صورت نیاز) و تفسیر نتایج خواهد بود.

از سوی دیگر، این پروژه با درک چالش‌های ورود به پروژه‌های واقعی، به دنبال ایجاد فضایی امن و آموزشی برای افرادی است که هنوز در ابتدای مسیر حرفه‌ای خود قرار دارند و نیازمند تقویت اعتماد به نفس هستند. محوریت کار در این پروژه بر دو بخش اساسی استوار است: **تحلیل داده** به عنوان سنگ بنای کشف دانش از دل اطلاعات، و آشنایی با رویکردها و مفاهیم کلیدی در **پروژه‌های علم داده** که می‌توانند در پروژه‌های پیچیده‌تر مورد استفاده قرار گیرند.

با همراهی در این سفر تحلیلی، نه تنها مهارت‌های فنی خود را ارتقا خواهید داد، بلکه با چالش‌های واقعی داده‌های دنیای کسب‌وکار نیز آشنا خواهید شد و گامی محکم‌تر به سوی تبدیل شدن به یک متخصص داده برخواهید داشت.

استفاده از این پروژه برای مصارف تجاری ممنوع است و باید از صاحب اثر اجازه گرفته شود.

* استخراج داده ها از هاگین فیس
* انتقال داده ها به SQL Server
* تجزیه و تحلیل اولیه داده ها در SQL
* وصل کردن داده های درج شده در SQL به پایتون
* EDA پایتون
* انتخاب ویژگی
* تولید ویژگی (ایجاد ویژگی های جدید)
* گزارش اول برای تمام داده ها
* پاکسازی داده ها
* ایجاد مدل ها
* نمایش نمودارها در Power BI
* نصب Power BI
* وارد کردن داده ها به عنوان SQL Servers
* ایجاد داشبورد جدید
* نوشتن گزارش
* ارائه
* مهلت: 1 خرداد 1404

**۱. تعریف مسئله و اهداف**

* **هدف کلی:** به دست آوردن بینش ارزشمند از داده های تبلیغات املاک برای پاسخ به سوالات تجاری کلیدی و کشف الگوها.
* **سوالات خاص:**
  + چه عواملی بیشترین تأثیر را بر قیمت ملک دارند؟
  + آیا می توانیم قیمت ملک را بر اساس ویژگی های آن پیش بینی کنیم؟
  + چه نوع املاکی در مناطق مختلف محبوب تر هستند؟
  + آیا الگوهای خاصی در امکانات و ویژگی های ملک با قیمت یا نوع ملک مرتبط وجود دارد؟
  + چگونه می توانیم تبلیغات موثرتری برای انواع مختلف ملک ایجاد کنیم؟

**۲. جمع آوری و درک داده ها**

* **منبع داده:** مجموعه داده تبلیغات املاک از [Hugging Face](https://huggingface.co/datasets/divaroffical/real_estate_ads).
* **اکتشاف اولیه داده ها (EDA):**
  + بررسی ساختار داده ها (ستون ها، انواع داده ها).
  + محاسبه آمارهای توصیفی (میانگین، میانه، انحراف معیار و غیره) برای ویژگی های عددی.
  + بررسی توزیع مقادیر برای ویژگی های دسته ای (تعداد هر دسته).
  + شناسایی مقادیر از دست رفته و نحوه توزیع آنها.
  + شناسایی داده های پرت احتمالی در ویژگی های عددی.
  + تجسم داده ها:
    - هیستوگرام ها برای نمایش توزیع ویژگی های عددی.
    - نمودارهای میله ای برای نمایش فراوانی دسته های مختلف در ویژگی های دسته ای.
    - نمودارهای جعبه ای برای شناسایی داده های پرت و مقایسه توزیع بین دسته ها.
    - نمودارهای پراکندگی برای بررسی روابط بین دو ویژگی عددی.
    - نقشه های جغرافیایی برای تجسم توزیع املاک در مناطق مختلف (اگر اطلاعات مکانی دقیق موجود باشد).

**۳. آماده سازی داده ها**

* **پاکسازی داده ها:**
  + رسیدگی به مقادیر از دست رفته (حذف، انتساب با میانگین/میانه/مد، یا روش های پیشرفته تر).
  + حذف داده های تکراری.
  + اصلاح داده های ناسازگار یا اشتباه.
  + مدیریت داده های پرت (حذف یا تبدیل).
* **مهندسی ویژگی:**
  + ایجاد ویژگی های جدید از ویژگی های موجود که ممکن است برای مدل سازی مفید باشند (به عنوان مثال، نسبت قیمت به مساحت، تعداد امکانات).
  + تبدیل ویژگی های دسته ای به فرمت عددی قابل استفاده برای مدل ها (به عنوان مثال، استفاده از One-Hot Encoding یا Label Encoding).
  + پردازش ویژگی های متنی (اگر وجود داشته باشد) با استفاده از تکنیک های NLP مانند TF-IDF یا Word Embeddings.
  + مقیاس بندی ویژگی های عددی برای جلوگیری از تأثیر زیاد ویژگی هایی با مقادیر بزرگتر بر مدل ها (به عنوان مثال، Standardization یا Normalization).
* **انتخاب ویژگی:**
  + شناسایی ویژگی های مرتبط با اهداف تحلیلی با استفاده از روش های آماری (به عنوان مثال، ضریب همبستگی) یا روش های مبتنی بر مدل.
  + کاهش ابعاد داده ها در صورت لزوم (به عنوان مثال، با استفاده از PCA) برای بهبود عملکرد مدل و کاهش پیچیدگی.

**۴. تحلیل و مدل سازی (بسته به اهداف خاص)**

* **تحلیل توصیفی:**
  + خلاصه کردن ویژگی های کلیدی برای درک بهتر مجموعه داده.
  + شناسایی روندها و الگوها در داده ها (به عنوان مثال، افزایش قیمت در مناطق خاص).
  + مقایسه ویژگی ها بین گروه های مختلف (به عنوان مثال، مقایسه میانگین قیمت آپارتمان ها و ویلاها).
* **مدل سازی پیش بینی (اگر هدف پیش بینی قیمت است):**
  + انتخاب مدل های رگرسیون مناسب (به عنوان مثال، رگرسیون خطی، رگرسیون چند جمله ای، درخت های تصمیم گیری، جنگل های تصادفی، گرادیان بوستینگ).
  + تقسیم داده ها به مجموعه های آموزشی و آزمایشی.
  + آموزش مدل بر روی مجموعه آموزشی.
  + ارزیابی عملکرد مدل بر روی مجموعه آزمایشی با استفاده از معیارهای مناسب (به عنوان مثال، میانگین خطای مربع، R-squared).
  + تنظیم پارامترهای مدل برای بهبود عملکرد.
* **مدل سازی طبقه بندی (اگر هدف طبقه بندی نوع ملک یا ویژگی دیگر است):**
  + انتخاب مدل های طبقه بندی مناسب (به عنوان مثال، رگرسیون لجستیک، ماشین های بردار پشتیبان، درخت های تصمیم گیری، جنگل های تصادفی، شبکه های عصبی).
  + تقسیم داده ها به مجموعه های آموزشی و آزمایشی.
  + آموزش مدل بر روی مجموعه آموزشی.
  + ارزیابی عملکرد مدل بر روی مجموعه آزمایشی با استفاده از معیارهای مناسب (به عنوان مثال، دقت، فراخوان، F1-score).
  + تنظیم پارامترهای مدل برای بهبود عملکرد.
* **تحلیل خوشه بندی (اگر هدف یافتن گروه های مشابه از املاک است):**
  + انتخاب الگوریتم های خوشه بندی مناسب (به عنوان مثال، K-Means، DBSCAN).
  + اعمال الگوریتم بر روی داده های آماده شده.
  + تجزیه و تحلیل ویژگی های خوشه های به دست آمده.

**۵. ارزیابی و تفسیر نتایج**

* ارزیابی عملکرد مدل های ساخته شده با استفاده از معیارهای مناسب.
* تفسیر نتایج مدل ها و تحلیل ها در چارچوب سوالات تجاری تعریف شده.
* شناسایی بینش های کلیدی و الگوهای کشف شده در داده ها.
* ارائه نتایج به صورت واضح و قابل فهم با استفاده از تجسم ها و خلاصه های مناسب.

**۶. مستندسازی و ارائه**

* مستندسازی کامل مراحل انجام شده، کدها و نتایج به دست آمده.
* تهیه یک گزارش یا ارائه جامع برای به اشتراک گذاشتن یافته ها و بینش ها با ذینفعان.
* ارائه پیشنهادات و توصیه های مبتنی بر نتایج تحلیل ها.

**. تعریف مسئله و اهداف (به روز شده)**

* **هدف کلی:** به دست آوردن بینش ارزشمند از داده های تبلیغات املاک برای پاسخ به سوالات تجاری کلیدی، کشف الگوها و ارائه راهکارهایی برای ذینفعان.
* **سوالات خاص:**
  + **پیش‌بینی قیمت مسکن:** آیا می‌توان مدلی ساخت که قیمت مسکن را بر اساس ویژگی‌ها پیش‌بینی کند؟
  + **تحلیل روند بازار:** تغییرات قیمت و تقاضا در مناطق مختلف چگونه بوده و چه روندهایی در حال شکل‌گیری است؟
  + **شناسایی عوامل موثر بر قیمت:** کدام ویژگی‌ها (مانند متراژ، امکانات، موقعیت) بیشترین تأثیر را بر قیمت مسکن دارند؟
  + **تحلیل داده‌های اجاره:** وضعیت بازار اجاره چگونه است؟ چه عواملی بر قیمت اجاره تأثیر می‌گذارند؟
  + **مقایسه مناطق مختلف:** قیمت‌ها و تقاضا در شهرهای مختلف یا مناطق درون یک شهر چگونه با هم مقایسه می‌شوند؟
  + آیا می‌توان الگوهای خاصی در ارتباط بین ویژگی‌های ملک و قیمت/اجاره در مناطق مختلف شناسایی کرد؟
  + چگونه می‌توان از این تحلیل‌ها برای تصمیم‌گیری‌های بهتر در زمینه خرید، فروش و اجاره ملک استفاده کرد؟

**۲. جمع آوری و درک داده ها**

* **منبع داده:** مجموعه داده تبلیغات املاک از [Hugging Face](https://huggingface.co/datasets/divaroffical/real_estate_ads).
* **اکتشاف اولیه داده ها (EDA):** (همانند قبل با تأکید بیشتر بر ویژگی‌های مرتبط با اهداف جدید)
  + بررسی توزیع قیمت و اجاره.
  + بررسی توزیع املاک در مناطق مختلف.
  + بررسی ویژگی‌های مختلف املاک و ارتباط آنها با قیمت و اجاره.
  + تجسم روند قیمت و تقاضا در طول زمان (اگر داده‌های تاریخی موجود باشد).
  + مقایسه توزیع ویژگی‌ها و قیمت/اجاره در مناطق مختلف.

**۳. آماده سازی داده ها**

* **پاکسازی داده ها:** (همانند قبل با توجه به ویژگی‌های جدید)
  + اطمینان از صحت و یکپارچگی داده‌های مربوط به قیمت، اجاره و موقعیت.
* **مهندسی ویژگی:** (همانند قبل با تمرکز بر ویژگی‌های مرتبط با اهداف جدید)
  + ایجاد ویژگی‌های تعاملی بین موقعیت و سایر ویژگی‌ها (به عنوان مثال، میانگین قیمت در یک منطقه).
  + استخراج ویژگی‌های مرتبط با روند بازار (اگر داده‌های تاریخی موجود باشد، مانند تغییرات قیمت نسبت به زمان).
* **انتخاب ویژگی:** (همانند قبل با در نظر گرفتن تأثیر ویژگی‌ها بر قیمت، اجاره و تقاضا)

**۴. تحلیل و مدل سازی (با توجه به اهداف جدید)**

* **پیش‌بینی قیمت مسکن:**
  + استفاده از مدل‌های رگرسیون (همانند قبل).
  + ارزیابی عملکرد مدل با توجه به دقت پیش‌بینی قیمت.
  + شناسایی مهم‌ترین عوامل موثر بر قیمت از طریق ضرایب یا اهمیت ویژگی‌های مدل.
* **تحلیل روند بازار:**
  + استفاده از تحلیل سری‌های زمانی (اگر داده‌های تاریخی موجود باشد) برای شناسایی روندها و فصلیت در قیمت و تقاضا.
  + تجسم تغییرات قیمت و حجم معاملات در مناطق مختلف در طول زمان.
  + بررسی شاخص‌های بازار مسکن (اگر قابل محاسبه باشد).
* **شناسایی عوامل موثر بر قیمت:**
  + استفاده از تحلیل همبستگی و اهمیت ویژگی‌های مدل‌های پیش‌بینی قیمت برای تعیین قوی‌ترین عوامل.
  + انجام تحلیل‌های آماری برای بررسی تأثیرات فردی و ترکیبی ویژگی‌ها بر قیمت.
* **تحلیل داده‌های اجاره:**
  + انجام EDA جداگانه بر روی داده‌های اجاره (در صورت وجود).
  + ساخت مدل‌های پیش‌بینی اجاره (مشابه پیش‌بینی قیمت).
  + شناسایی عوامل موثر بر قیمت اجاره.
  + مقایسه روندها و عوامل موثر بین بازار خرید و اجاره.
* **مقایسه مناطق مختلف:**
  + مقایسه آمارهای توصیفی (میانگین، میانه قیمت/اجاره، حجم معاملات) بین مناطق مختلف.
  + استفاده از تجسم‌ها (مانند نمودارهای جعبه‌ای، نقشه‌های حرارتی) برای مقایسه توزیع قیمت و تقاضا در مناطق مختلف.
  + انجام آزمون‌های آماری برای بررسی تفاوت‌های معنادار بین مناطق.

**۵. ارزیابی و تفسیر نتایج**

* ارزیابی عملکرد مدل‌های پیش‌بینی قیمت و اجاره.
* تفسیر نتایج تحلیل روند بازار و شناسایی روندهای کلیدی.
* تعیین میزان تأثیرگذاری هر یک از ویژگی‌ها بر قیمت و اجاره.
* مقایسه و ارزیابی تفاوت‌های بازار در مناطق مختلف.
* ارائه بینش‌های عملی برای تصمیم‌گیری‌های مربوط به خرید، فروش و اجاره ملک.

**۶. مستندسازی و ارائه**

* مستندسازی کامل مراحل انجام شده و نتایج به دست آمده.
* ارائه یافته‌ها به صورت واضح و جذاب، با تأکید بر پاسخ دادن به سوالات اولیه و ارائه پیشنهادات مشخص.
* استفاده از تجسم‌های مناسب برای انتقال موثر بینش‌ها.