

به نام خدا

« پروژه اول - فاز ۱ »

بوتکمپ علم داده کوئرا

تابستان ۱۴۰۲



آخرین ویرایش: چهارشنبه ۸ شهریور

مقدمه و آشنایی



به فاز نخست از پروژه‌ی اول بوت‌کمپ دیتا ساینس خوش آمدید. این فاز دارای چهار مرحله اصلی با تمرکز بر جمع‌آوری، ذخیره‌سازی، تحلیل داده و داشبور딩 می‌باشد.

- در مرحله‌ی اول، به جمع‌آوری دیتا از سطح اینترنت با استفاده از Web Scraping می‌پردازیم.
- در مرحله‌ی دوم، به ذخیره‌سازی دیتای استخراج‌شده در دیتابیس می‌پردازیم. در این قسمت نیاز است درباره‌ی چگونگی انجام این کار به بهترین شکل ممکن، ایده‌پردازی شود.
- در مرحله‌ی سوم، به دنیای آمار سفر کرده و با استفاده از آزمون‌ها و نمودارهای آماری، به بررسی و شناخت بیشتر داده‌های پروژه می‌پردازیم.
- در مرحله‌ی آخر، با استفاده از ابزارهای مصورسازی‌ای که Power BI در اختیارمان می‌گذارد به کشف بینش‌ها (insights) و الگوهایی که ممکن است در نگاه اول قابل تشخیص نباشند پرداخته و یک داشبورد تحلیلی جذاب ساخته خواهد شد.

توجه داشته باشید که در کنار مهارت‌های سخت، بخش زیادی از نحوه‌ی پیش‌برد پروژه به مهارت‌های نرم شما مربوط است. بنابراین در طول پروژه نیاز خواهید داشت تا با کار تیمی و تعامل و پشتیبانی از یکدیگر، به بهترین نحوه ممکن با هم‌تیمی‌های خود در ارتباط باشید و به پیشرفت پروژه کمک کنید. مشارکت شما به‌عنوان عضوی از تیم توسط منتور گروه مورد ارزیابی قرار خواهد گرفت.

ممکن است بخش‌های مختلف پروژه وابسته به همدیگر باشند، بنابراین در اجرای هر بخش سعی کنید نیازمندی‌های بخش‌های دیگر را مطالعه کرده و تامین کنید.

بخش اول: استخراج داده

در این پروژه به کار بر روی اطلاعات مرتبط با رمزارزها پرداخته خواهد شد و بدین منظور از داده‌های موجود در سایت [CoinMarketCap](https://coinmarketcap.com) استفاده خواهیم کرد. از آن‌جا که به دلیل دینامیک بودن صفحه‌ی نخست این وبسایت ممکن است استخراج داده‌ها از این صفحه با دشواری‌های بیشتری همراه باشد، قصد داریم از بخش تاریخچه‌ی آن ([لینک](#)) استفاده کرده و اطلاعات رمزارزها را برای یک بازه‌ی زمانی مشخص استخراج کنیم.

در این صفحه می‌توانید با انتخاب یک تاریخ مشخص، اطلاعات رمزارزها در آن روز را مشاهده کنید. ابتدا از شما می‌خواهیم لیست ۲۰۰ رمز ارز برتر (۲۰۰ سطر ابتدایی جدول) در تاریخ **2023/08/25** را استخراج کرده و داده‌های آن را با فرمت جدولی به شکل زیر آماده کنید:

Rank	Name	Symbol	MainLink	HistoricalLink
1	Bitcoin	BTC	https://coinmarketcap.com/currencies/bitcoin/	https://coinmarketcap.com/currencies/bitcoin/historical-data/
...

منظور از MainLink، لینک صفحه‌ی ویژه‌ی آن ارز در سایت CoinMarketCap است. منظور از HistoricalLink نیز لینک تاریخچه‌ی آن ارز است که هم می‌تواند از همین صفحه و هم از صفحه‌ی ویژه‌ی خود ارز استخراج شود.

Historical Snapshot - 05 May 2013

Market Cap:

All

Price:

All

Volume (24h):







All

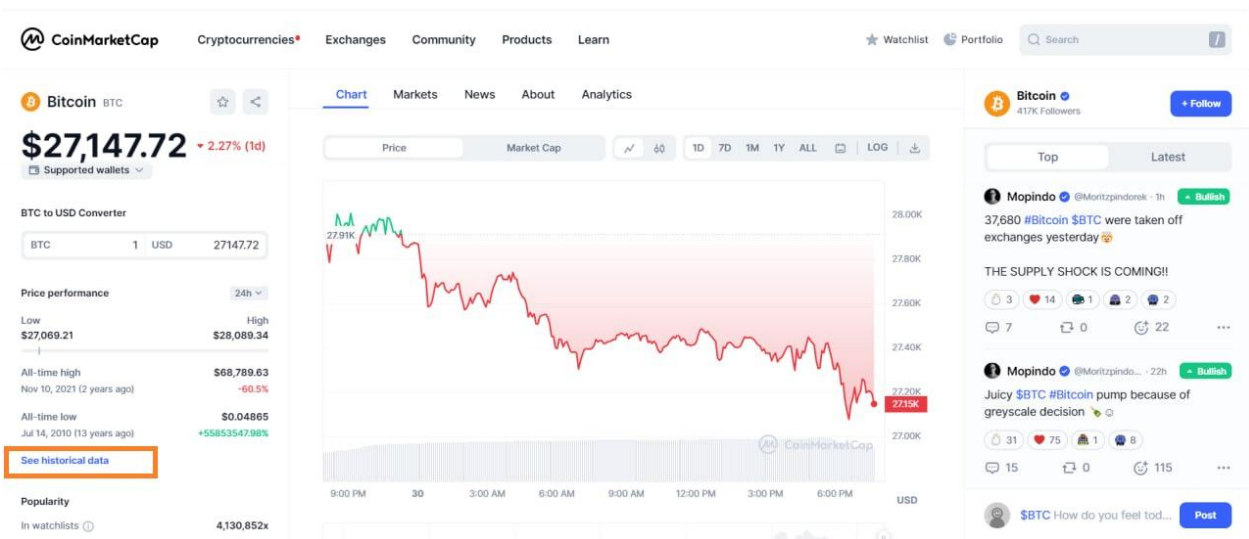
USD

Previous Week

Next Week

View All

Rank	Name	Symbol	Market Cap	Price	Circulating Supply	% 1h	% 24h	% 7d
1	 Bitcoin	BTC	\$1,288,693,216.22	\$115.91	11,118,050 BTC	0.43%	2.97%	-13.81%
2	 Litecoin	LTC	\$62,298,217.32	\$3.5909	17,348,954 LTC	0.10%	2.8	<div>View Chart</div> <div>View Markets</div> <div>View Historical Data</div>
3	 Namecoin	NMC	\$6,290,543.05	\$1.1510	5,465,350 NMC	-1.19%	8.9	-22.20%
4	 Peercoin	PPC	\$5,718,446.46	\$0.3037	18,830,240 PPC	0.31%	2.61%	-22.20%
5	 Feathercoin	FTC	\$2,017,436.23	\$0.313	6,444,650 FTC	-12.81%	-29.71%	-
6	 Freicoin	FRC	\$1,627,438.51	\$0.08114	20,057,908 FRC	-14.89%	-23.15%	-

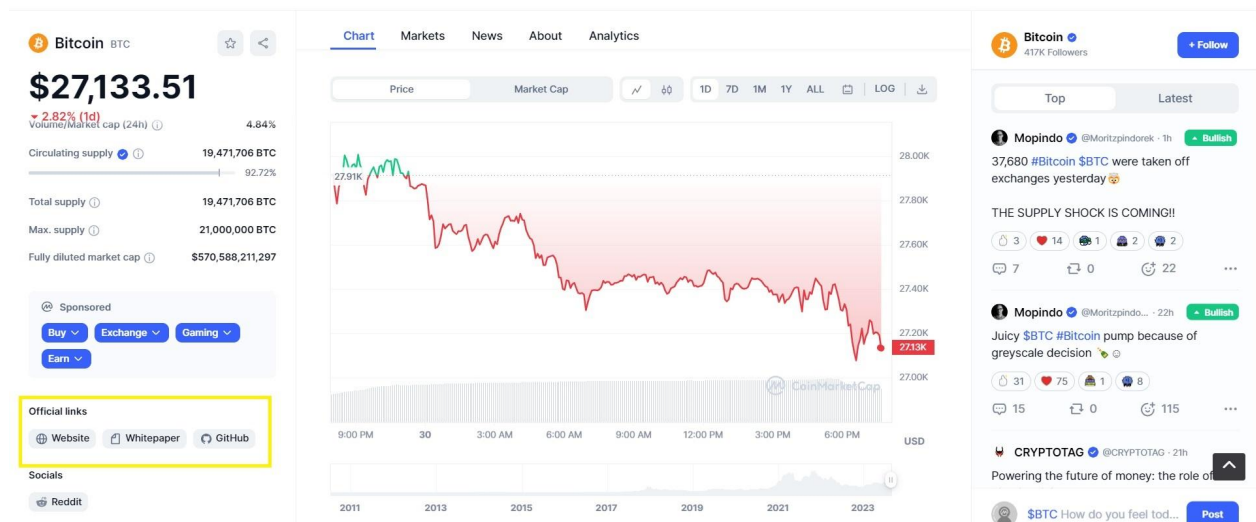


در صفحه‌ی تاریخچه‌ی هر ارز (مثل [این لینک](#)) می‌توان اطلاعات گوناگونی از جمله موارد زیر را مشاهده کرد:

- Name: نام رمزارز
- Symbol: نماد اختصاری رمزارز
- Market Cap: ارزش کل در گردش بازار ارز مربوطه
- Volume(24h): بیانگر این است که چقدر در ۲۴ ساعت اخیر این رمزارز در حال معامله شدن بوده
- Price: قیمت ارز
- Circulating supply: تعداد ارزی که در بازار به‌صورت عمومی در گردش بوده
- High: بالاترین ارزش پولی ارز در بازه‌ی زمانی مربوطه
- Low: پایین‌ترین ارزش پولی ارز در بازه‌ی زمانی مربوطه
- Open: ارزش پولی ارز در زمان شروع بازه‌ی زمانی معامله
- Close: ارزش پولی ارز در زمان پایان شروع بازه‌ی زمانی معامله

شما باید به‌ازای هر رمز ارز، اطلاعات ۳۶۵ روز اخیر آن را استخراج کنید. برای این‌کار می‌توانید از فایل‌های CSV که خود سایت در اختیارتان می‌گذارد استفاده کنید و تنها بازه‌ی تاریخی را معادل ۳۶۵ روز اخیر انتخاب کنید.

علاوه بر این، به‌ازای هر رمز ارز نیاز است اطلاعاتی را از صفحه‌ی ویژه‌ی آن استخراج کنید. یکی از این اطلاعات مهم، لینک صفحه‌ی گیت‌هاب آن (در صورت وجود) است که در قسمت Official links قابل مشاهده است.



مورد دیگر تگ‌هایی‌ست که در پایین همان بخش سمت چپ صفحه نوشته شده‌اند. به‌عنوان مثال برای بیت‌کوین سه تگ Mineable، PoW و SHA-256 نوشته شده است.

بخش دوم: طراحی دیتابیس

کاری که در این فاز از پروژه باید انجام شود، ساختن دیتابیس و جداول مورد نیاز و ایجاد ارتباط بین جدول‌هاست، در ادامه توضیحاتی را برای این قسمت آورده‌ایم:

1. طراحی دیتابیس:

- الف) موجودیت‌های اصلی در داده‌ها را تعیین کنید (مانند کوین‌ها، قیمت‌ها و غیره).
- ب) روابط بین موجودیت‌ها را تعیین کنید.
- پ) ویژگی‌های داده‌ای خاصی را که باید برای هر موجودیت ذخیره شود، تعیین کنید (مانند ویژگی‌های time_high و غیره).
- ت) از تکنیک‌های نرمالایز کردن برای حذف افزونگی داده‌ها و اطمینان از یکپارچگی داده‌ها استفاده کنید.

2. انتخاب ابزار و تکنولوژی دیتابیس:

یک سیستم مدیریت پایگاه داده (DBMS) را انتخاب کنید که به بهترین وجه با نیازهای شما مطابقت دارد (مانند MySQL). مقیاس پذیری و الزامات عملکرد برنامه را در نظر بگیرید تا مطمئن شوید که DBMS انتخابی می تواند حجم داده و حجم کاری مورد انتظار را مدیریت کند.

3. جمع آوری و ذخیره ی داده ها در دیتابیس:

با استفاده از ابزارهای Web Scraping و آموخته های خود داده های مورد نظر را استخراج نمایید و در دیتابیس خود ذخیره کنید.

انتظار می رود در هر جا که به پیش پردازش اضافی نیاز باشد طبق تصمیم خود این کار را انجام داده و دلایل منطقی ای برای آن داشته باشید. به عنوان مثال ممکن است به مدیریت داده های گم شده، تغییر مقیاس، نرمال سازی ویژگی ها، کدگذاری ویژگی های دسته ای و غیره نیاز پیدا کنید.

بخش سوم: تحلیل‌های آماری

در این بخش، به کمک دانش آماری می‌خواهیم به تعدادی از سوال‌ها پاسخ دهیم؛ برخی از سوالات برای درک و یافتن شهود از داده‌ها پرسیده شده است، برخی دیگر از سمت یک شخص خاص، و در انتها تعدادی فرضیه مطرح شده است که شما باید آن‌ها را اعتبارسنجی کنید.

آمار توصیفی

1. ارتباط میان ارزش بازار (Market Cap) و حجم معاملاتی روزانه (Volume 24h) رمزارزها را بررسی نمایید. برای این کار می‌توانید Scatter Plot آن را رسم کنید.
2. در یک سال اخیر کدام جفت‌رمزارزها بیشترین روزها با تغییر قیمت هم‌سو را تجربه کرده‌اند و چند روز؟ (منظور از تغییر قیمت هم‌سو افزایش قیمت هر دو رمز ارز یا کاهش قیمت جفت‌شان طی یک روز است) ۳۰ جفت رمزارز برتر را نخست به ترتیب تعداد روز هم‌سو، سپس نام رمز ارز اول و در نهایت نام رمز ارز دوم نمایش دهید؛ همچنین توجه داشته باشید همواره نام رمزارز اول از نظر حروف الفبا مقدم بر نام رمز ارز دوم باشد.
3. توزیع حجم معاملات روزانه رمزارزهای قابل استخراج (Mineable) را رسم کنید.
4. ماتریس هم‌بستگی را برای تغییرات قیمت ۱۶ رمزارز برتر از نظر ارزش بازار رسم نمایید.
5. روزهایی که بیش از ۳۵ درصد رمزارزهای مورد مطالعه افزایش قیمت را تجربه می‌کنند روزهای سبز و دیگر روزها را روزهای قرمز بازار می‌نامیم. ۱۰ رمزارز که بیش‌ترین تعداد افزایش قیمت را در روزهای قرمز در بهار ۲۰۲۳ (برج‌های ۳، ۴ و ۵ میلادی) تجربه کرده‌اند کدام رمزارزها هستند؟ حجم بازار این ۱۰ رمزارز را به صورت نمودار میله‌ای نمایش دهید.

تخمین

- به صورت تصادفی ۴۰ رمز ارز را از میان داده‌های استخراج شده انتخاب نمایید و میانگین حجم معاملاتی روزانه هر یک را به دست آورید. طبق این نمونه برداری بازه‌ی اطمینان ۹۸ درصد را برای حجم معاملاتی محاسبه نمایید.

آزمون فرض

- شرکت کوئرا شغل پاره‌وقتی را به شما پیشنهاد داده و ۲ حق انتخاب برای روزهای کاری به شما داده است. انتخاب اول، روزهای چهارشنبه تا شنبه و انتخاب دوم روزهای یکشنبه تا چهارشنبه می‌باشد. از آن‌جا که شما علاوه بر فعالیتهای کاری در حوزه‌ی تخصصی خود، به خرید و فروش رمز ارز نیز مشغول هستید، تصمیم می‌گیرید یک مطالعه‌ی آماری بر روند بازار رمزارزها داشته باشید تا روزهای پویا و متلاطم‌تر بازار را از دست ندهید. بدین منظور می‌خواهید میانگین میزان تغییر قیمت رمزارزها در روزهای پنج‌شنبه، جمعه و شنبه را با روزهای یکشنبه، دوشنبه و سه‌شنبه مقایسه کنید. آیا با این شاخص انتخاب شده، تفاوت فاحشی میان دو انتخاب ممکن برای روزهای کاری وجود دارد؟ در صورت مثبت بودن پاسخ، انتخاب شما چه روزهایی است؟
- معمولاً افرادی که تازه وارد بازار رمزارز می‌شوند، مایل به سرمایه‌گذاری در رمزارزهای بسیار معروف هستند و از طرفی دیگر برخی افراد ریسک‌پذیری پایینی دارند و تمایل دارند در رمزارزهایی با تغییرات کم‌تر نسبت به دیگر رمزارزها سرمایه‌گذاری کنند، لذا به نظر می‌رسد بیشتر معاملات بازار بین معروف‌ترین و باثبات‌ترین رمزارزهای بازار باشد. اکنون می‌خواهیم بررسی کنیم که چه میزان از معاملات روزانه بازار مربوط به این رمزارزهاست. برای این کار سه رمز ارز 'Bitcoin', 'Ethereum', 'Tether USDt' را در نظر بگیرید و بررسی کنید آیا این ادعای زیر صحیح است یا نه؟

"میانگین حجم معاملات روزانه‌ی 'Bitcoin'، 'Ethereum' و 'Tether USDt' به شدت بیشتر از

میانگین حجم معاملات روزانه باقی رمزارزهاست."

بخش چهارم: Power BI

در این بخش با استفاده از دیتابیزی که ساخته‌اید نمودارهای زیر را رسم کنید و یک داشبورد منطقی برای آن طراحی کنید.

1. توزیع حجم معاملات روزانه به تفکیک سال، ماه، هفته و روز
2. توزیع حجم معاملات روزانه به تفکیک روزهای هفته
3. نمودار Candlestick بر اساس تاریخ به همراه یک فیلتر برای انتخاب نام کوین و همچنین قابلیت Drill-Down روی تاریخ نمودار. اگر به صفحه یکی از کوین‌ها مراجعه کنید می‌توانید نمونه‌ی نموداری که مد نظر است را ببینید.
4. نموداری رسم کنید که در تاریخ دلخواه بتوانیم محاسبه کنیم که اگر یک کوین را در روز قبل با پایین‌ترین قیمت می‌خریدیم و در آن تاریخ مذکور (فردای روز خرید) با بالاترین قیمت می‌فروختیم چقدر سود بدست می‌آوردیم. برای رسم این نمودار ابتدا باید اختلاف میان بالاترین قیمت روز انتخاب شده با پایین‌ترین قیمت روز قبلش را برای کوین‌های مختلف محاسبه کنید در نهایت با مصورسازی آنها روی نمودار به مقایسه میزان سود هر کوین از این استراتژی بپردازید.

بخش‌های امتیازی

در طی مراحل پروژه، برای ایده‌های نوآورانه، راه‌حل‌های خلاقانه، تحلیل‌ها و نمودارهای معنادار دیگری که به آن‌ها اشاره نشده است، نمره‌ی امتیازی در نظر گرفته می‌شود. به‌طور مثال می‌توان به موارد زیر اشاره کرد:

- استفاده از دیتابیس در تمام مراحل که نیاز به خوانش داده‌ها (Power BI) و موارد استفاده از پایتون وجود دارد.
- جمع‌آوری داده‌های متنوعی از گیت‌هاب مرتبط با هر رمزارز همچون زبان‌های برنامه‌نویسی، اطلاعات مشارکت‌کنندگان، تعداد کامیت‌ها و غیره
- طراحی آزمون‌های آماری دیگر. به عنوان نمونه می‌توانید به حل سوال زیر بپردازید:

نوید ادعا می‌کند تغییر قیمت رمزارزها در روزهای تعطیل از یک توزیع نرمال پیروی می‌کند. بررسی کنید این ادعا تا چه اندازه صحیح است. بدین منظور نخست هیستوگرام تغییر قیمت روزانه رمزارزها در روزهای تعطیل (شنبه و یکشنبه) را رسم نمایید. سپس از آزمون آماری متناسب با مسئله استفاده کرده و نظر خود را در رابطه با این ادعا بیان کنید.

- اطلاعات و نمودارهای مفید بیشتر در داشبورد
- مستندسازی دقیق، استفاده از گیت، کدنویسی تمیز و استفاده از OOP و غیره

موفق باشید 😊