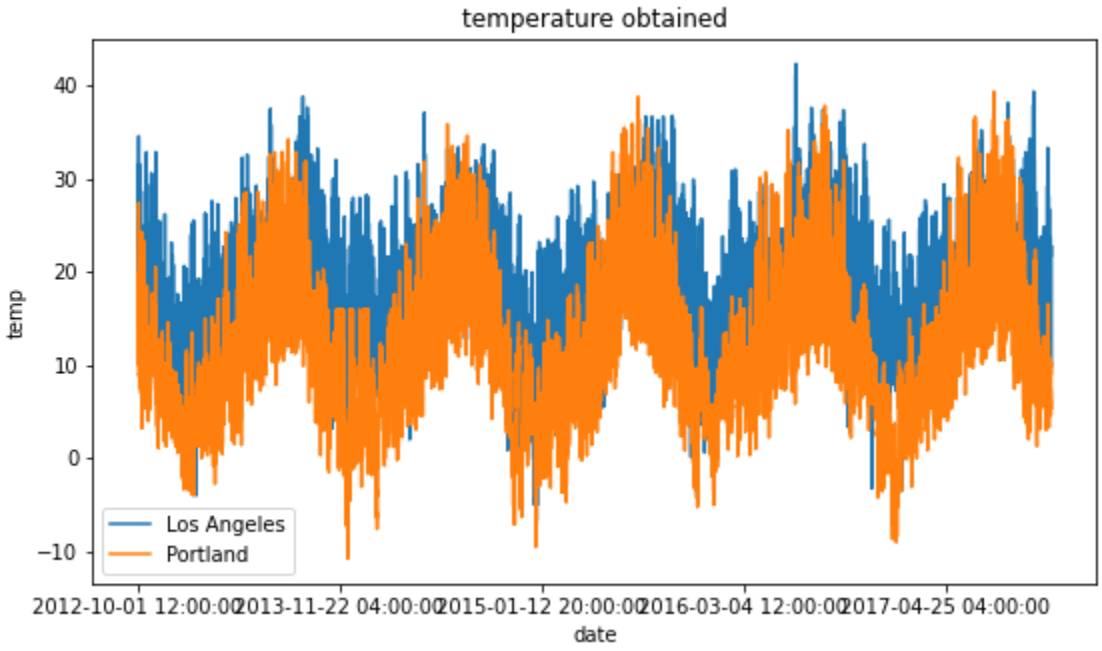
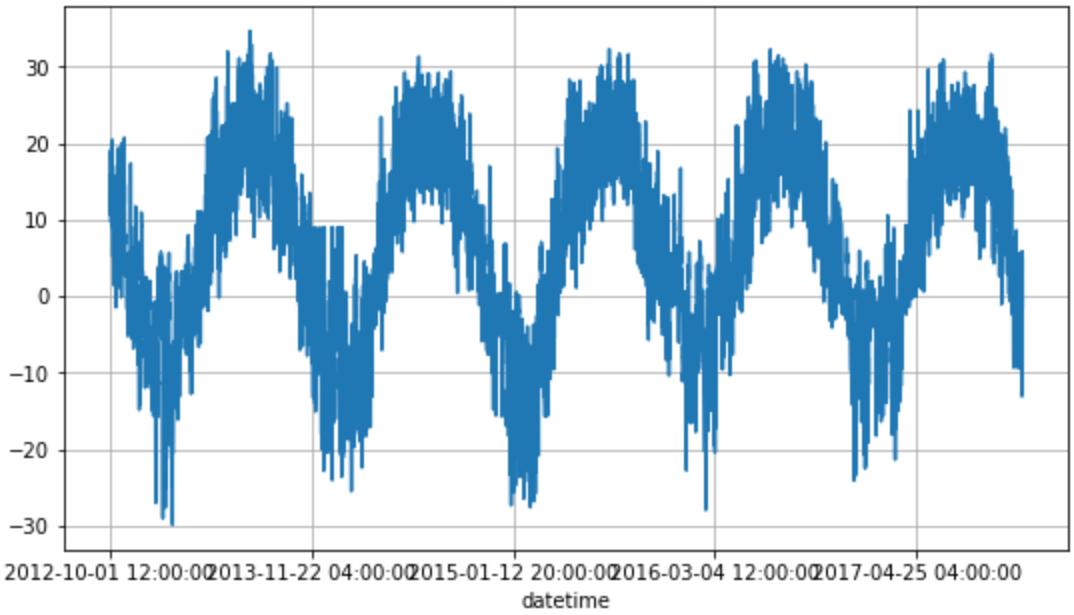
گزارش تمرین امتیازی علوم داده

محمدرضا صیدگر-97222055

پروژه راجع به کار با داده های سری زمانی است و پیشبینی این نوع از داده ها که تفاوت این نوع داده ها در این است این داده ها به صورت دنباله ای هستند و تربیت آن ها اهمیت دارد ولی در پروژه های قبلی داده ها از نظر ترتیبی اهمیتی نداشتند و این ترتیب در این داده ها بر اساس زمان است.
کار ما در این پروژه اول بررسی trend داده ها سپس seasonality و stationarity است و درنهایت بررسی مدل های مختلف پیشبینی روی این نوع از داده ها.
من در این پروژه داده های دمای شهر های پورتلند ، لس آنجلس و مونترال را گرفتم و موارد فوق را برای هرکدام بررسی کردم (مدل ها فقط روی شهر پورتلند).
قبل از اینکه بدانیم مفاهیم فوق چی هستند در رابطه با پیش پردازش داده ها باید گفت که شهر هایی انتخاب شدند که داده های null کمی داشتند اما همان تعداد کم هم حذف نشدند و با این روش پر شدند میانگین داده قبلی و بعدی به جای آن داده null قرار گرفت. داده های اصلی یکای کلوین دارند ولی به سانتیگراد تبدیل شد که از نظر مفهومی هم قابل درک باشند.
در شکل زیر تغییرات دمای شهرهای پورتلند و لس آنجلس را در سال های 2012 تا 2017 می بینیم:



شکل زیر هم تغییرات دمای شهر مونترال است:

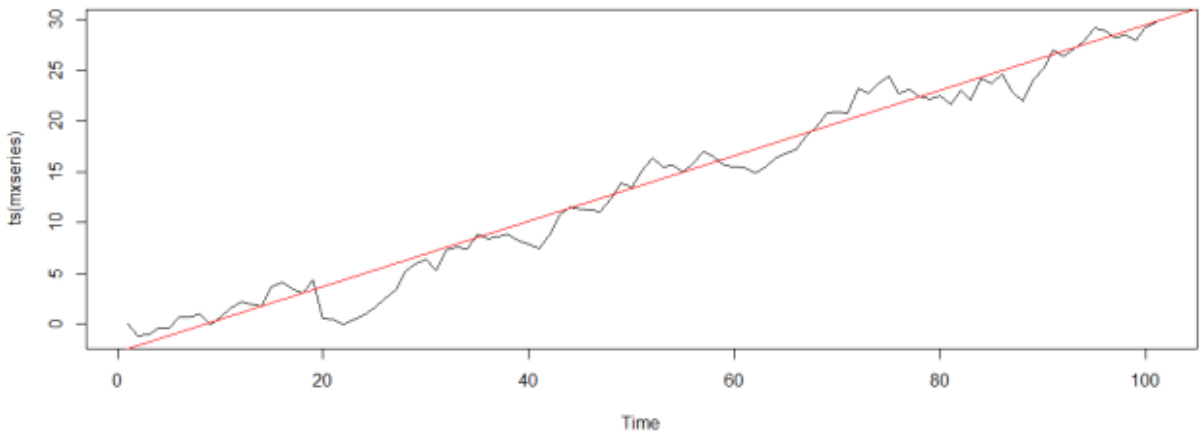


که همانطور که از نمودار ها هم مشخص است به طور کل دمای شهر لس آنجلس از پورتلند بیشتر و دمای پورتلند از مونترال بیشتر است.

حال 3 پرسش مطرح می شود:

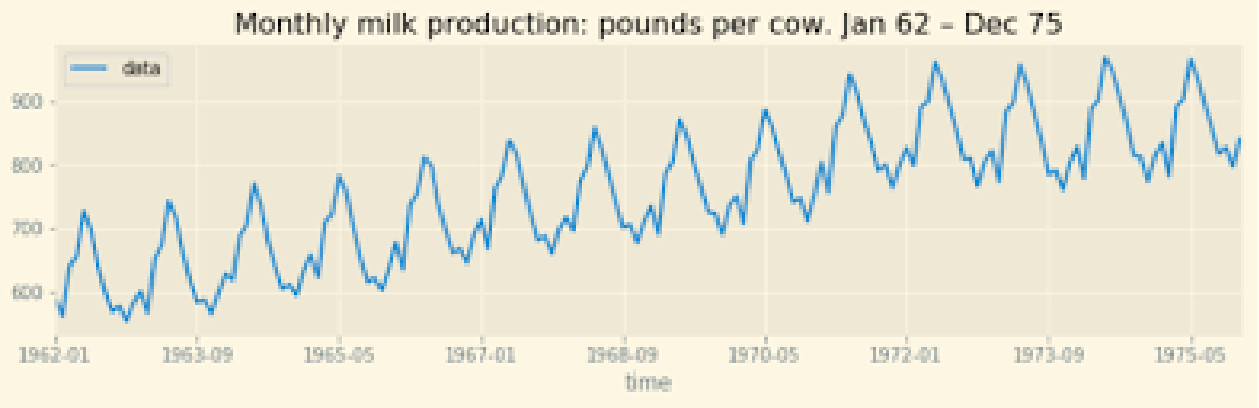
پرسش اول مطرح شده این است trend که چیست؟

trend الگویی در داده ها است که حرکت یک سری را به مقادیر نسبتاً بالاتر یا کمتر در یک دوره زمانی طولانی نشان می دهد. به عبارت دیگر، زمانی که یک شیب افزایشی یا کاهشی در سری زمانی وجود داشته باشد، trend مشاهده می شود. trend معمولاً برای مدتی اتفاق می افتد و سپس ناپدید می شود، تکرار نمی شود.



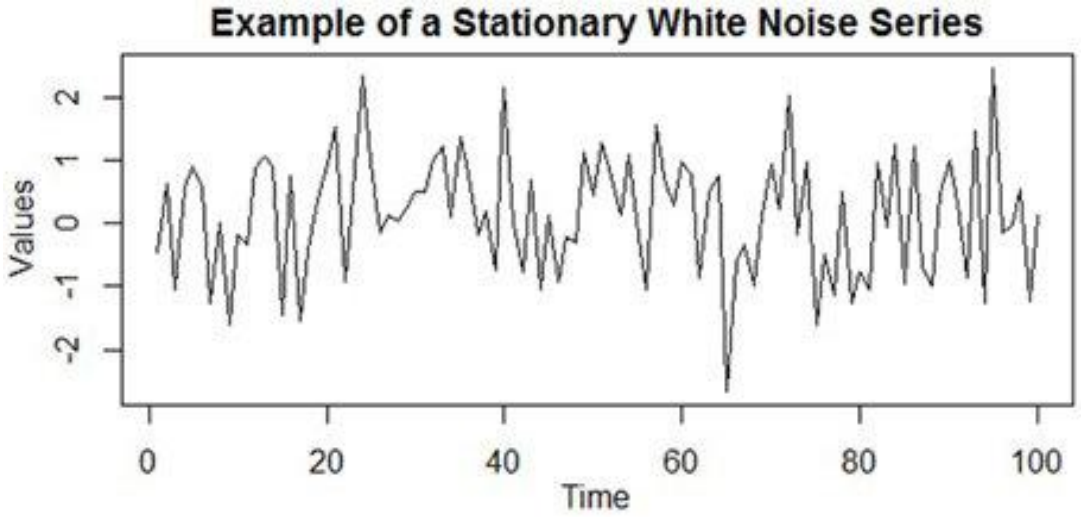
پرسش دوم این است که seasonality چیست؟

seasonality مشخصه یک سری زمانی است که در آن داده ها تغییرات منظم و قابل پیش بینی را تجربه می کنند که در هر سال تقویمی تکرار می شود. هر گونه نوسان یا الگوی قابل پیش بینی که در یک دوره یک ساله تکرار یا تکرار شود فصلی است.



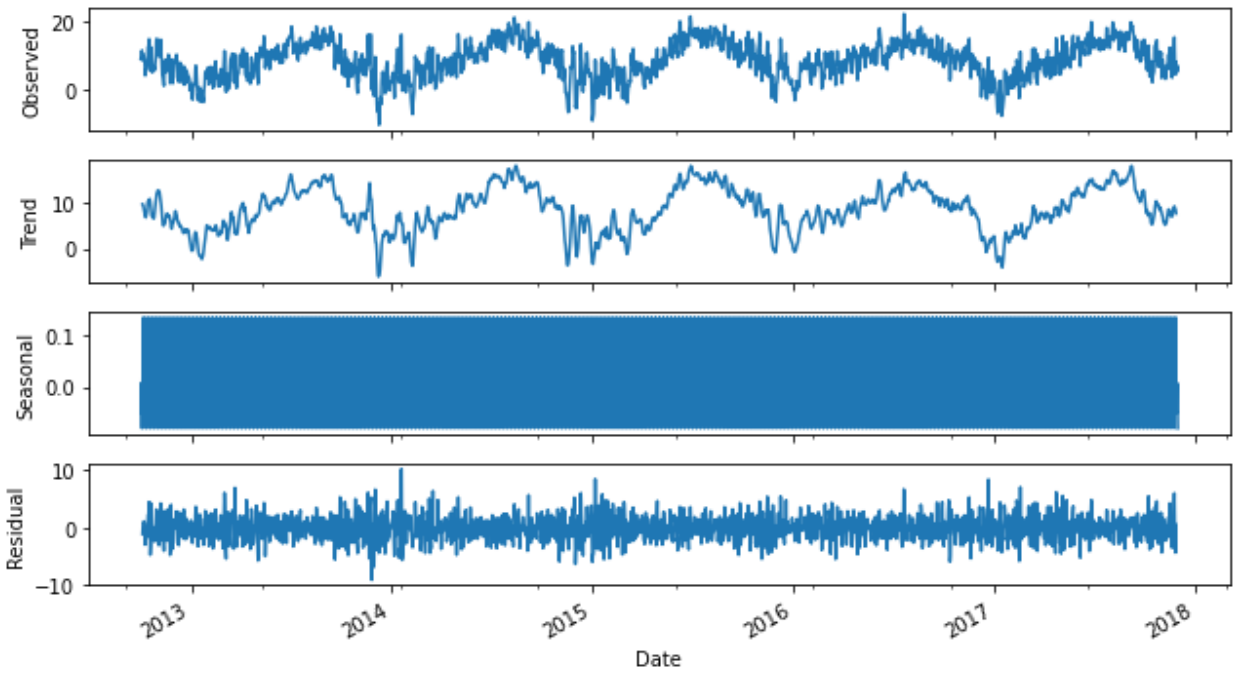
پرسش سوم این است که stationarity چیست؟

در شهودی ترین مفهوم، stationarity به این معناست که ویژگی‌های آماری فرآیندی که یک سری زمانی تولید می‌کند در طول زمان تغییر نمی‌کند. این بدان معنا نیست که سری در طول زمان تغییر نمی کند، فقط نحوه تغییر آن در طول زمان تغییر نمی کند.



حال که با این مفاهیم آشنا شدیم trend و seasonality را روی داده های این شهر ها مشخص کردیم و همینطور با آزمون ADF statistic پایداری آن ها را بررسی کردیم:

نمودار های زیر مربوط به شهر پورتلند است:

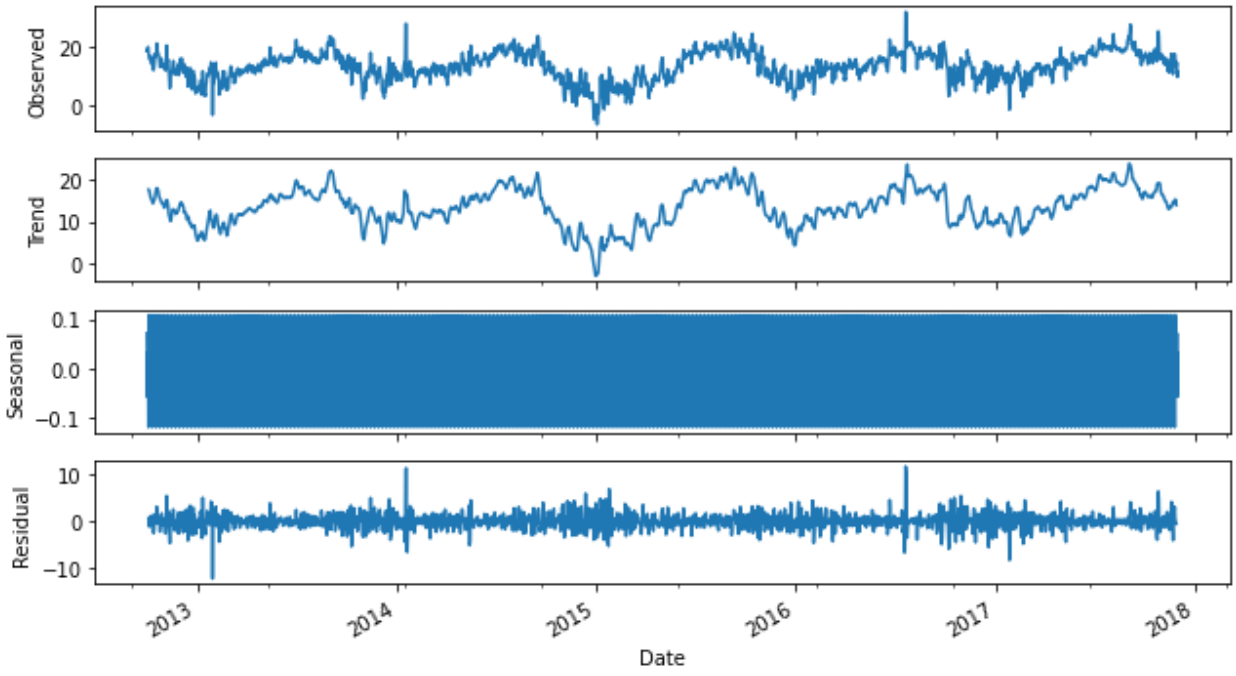


ADF Statistic: -3.854065

p-value: 0.002400

که نشان میدهد داده های شهر پورتلند پایدار هستند.

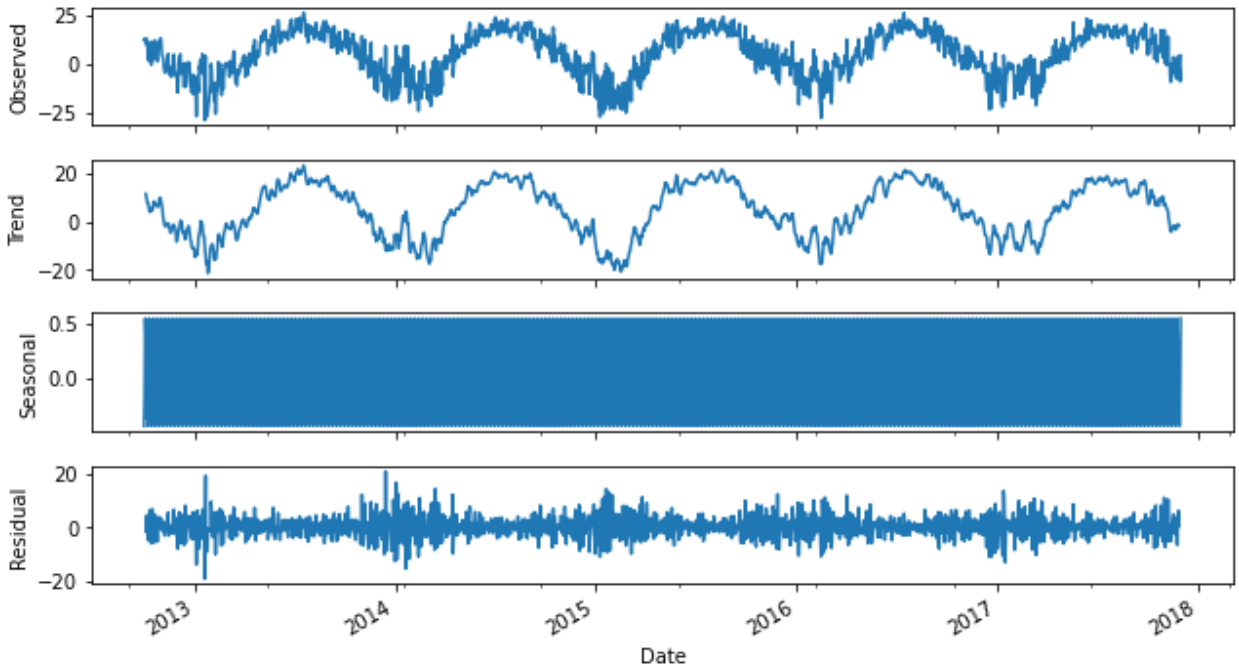
نمودار های زیر مربوط به شهر لس آنجلس است:



ADF Statistic: -3.805801

p-value: 0.002849
نشان میدهد داده های شهر لس آنجلس پایدار هستند.

نمودار های زیر مربوط به شهر مونترال است:



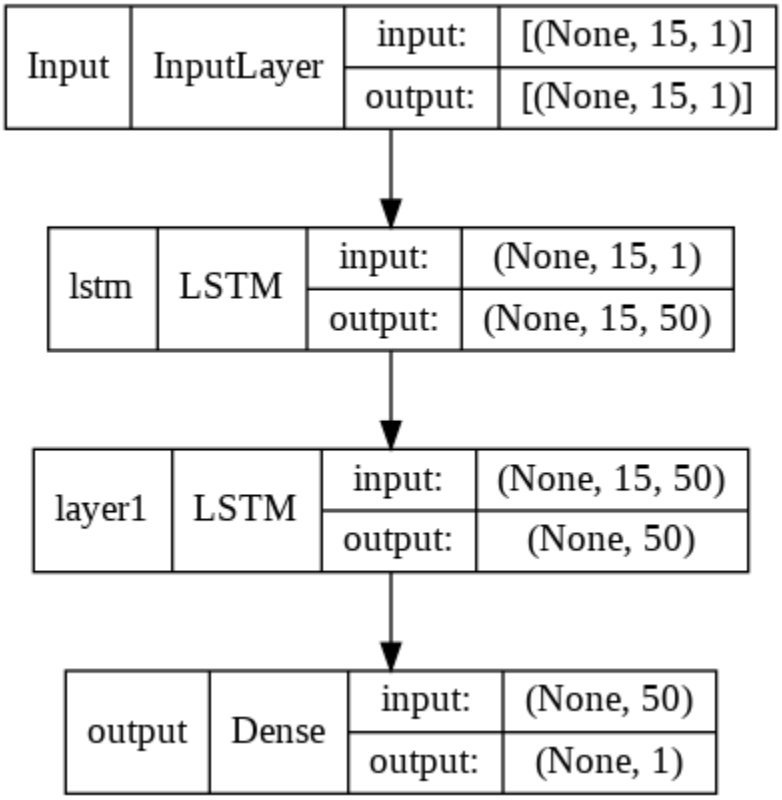
ADF Statistic: -2.504755
p-value: 0.114322

اما چون برای شهر مونترال p_value بیشتر از 0.05 شد پس داده های این شهر آنقدر پایدار نیستند.

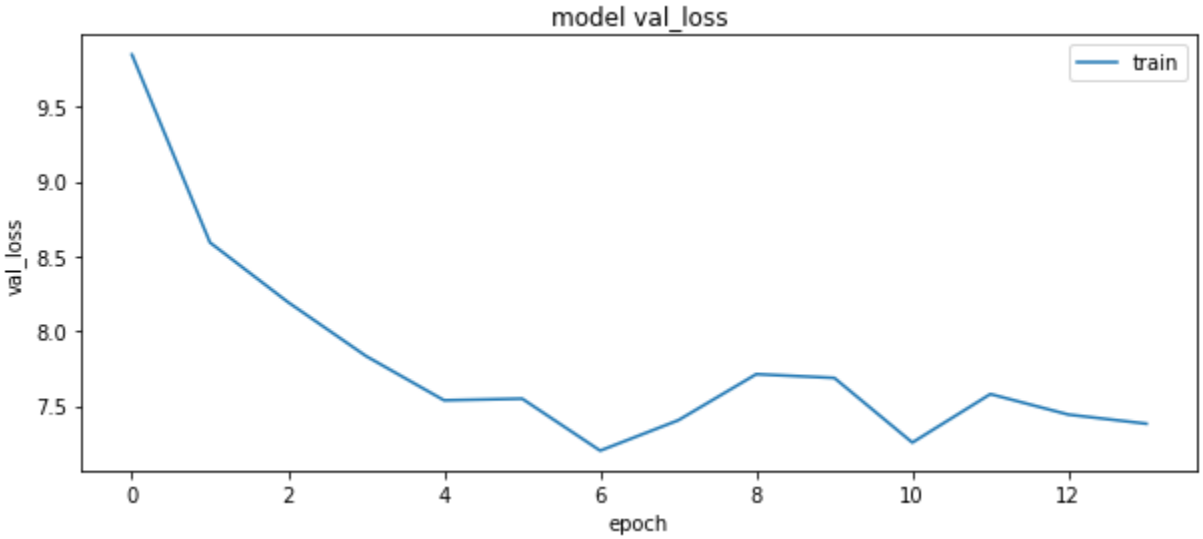
حال برای اینکه پیشبینی (forecasting) کنیم از مدل های مختلف مناسب داده های سری زمانی استفاده کردیم که نتایج و توضیحات هر کدام را در زیر خواهیم دید:

1)شبکه های عصبی RNN:
شبکه‌های عصبی بازگشتی (RNN) مدل‌های یادگیری عمیق هستند که معمولاً برای حل مشکلات داده‌های ورودی متوالی مانند سری‌های زمانی استفاده می‌شوند. RNN ها نوعی شبکه عصبی هستند که حافظه ای از آنچه قبلاً پردازش کرده است را حفظ می کند و بنابراین می تواند از تکرارهای قبلی در طول آموزش یاد بگیرد. حال برای اینکه داده ها را به این شبکه بدیم با ورودی 15 نورون باید به طور مثال برای اولین داده ای که دریافت میکند دمای 15 روز متوالی را بگیرد و به عنوان خروجی دمای روز 16 ام را پیشبینی کند.
ما از شبکه LSTM استفاده کردیم که یکی از انواع شبکه های RNN است .

معماری شبکه ما به شکل زیر است:



به عنوان ورودی 15 نورون دارد و سپس 2 لایه LSTM با 50 نورون و در نهایت خروجی Dense که 1 نورون است چون مسئله ما رگرسیون است و باید یک عدد پیوسته را پیشبینی کند.
از optimizer آدام استفاده شد و loss هم mse در نظر گرفتیم که نتایج به شکل زیر شد:



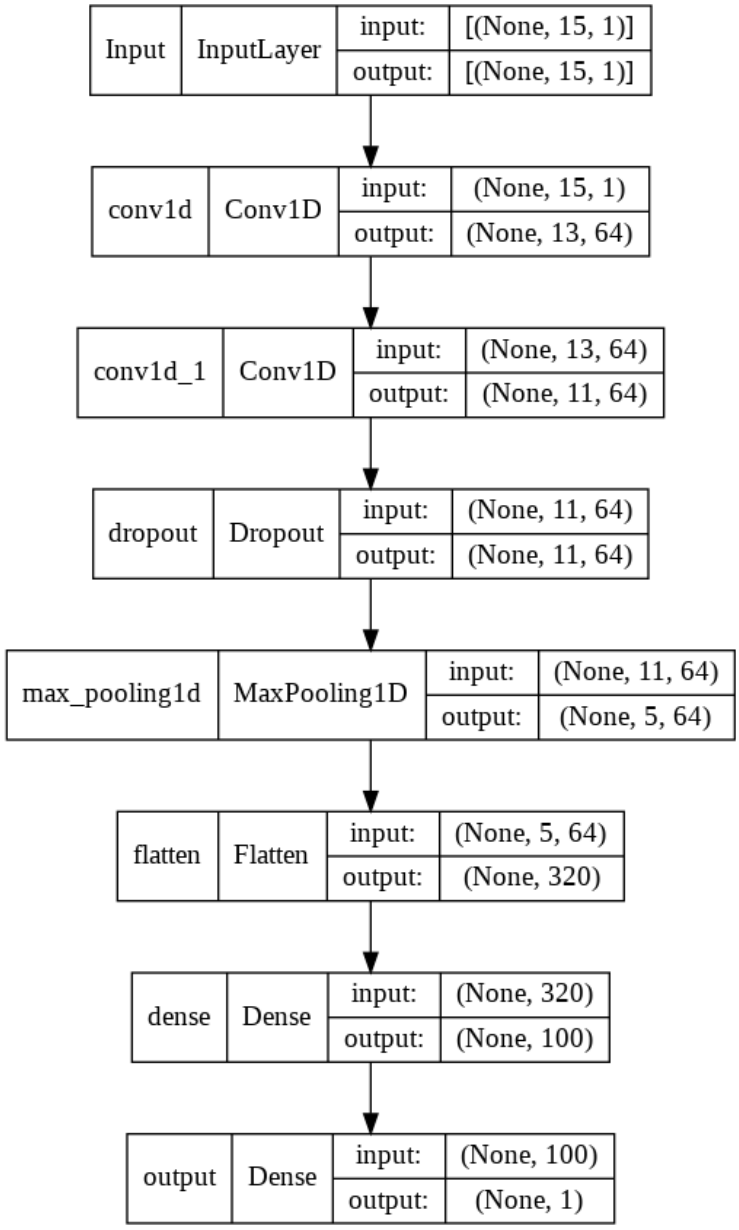
همینطور loss نهایی روی داده های train 7.04 شد و روی داده های validation 7.38 شد.

2) شبکه های عصبی Convolutional:

معمولا شبکه های CNN در پردازش تصویر استفاده می شوند اما از طرفی از Conv 1d در داده های سری زمانی هم استفاده می شود که کمک میکند به پیدا کردن pattern ها در این نوع از داده ها.

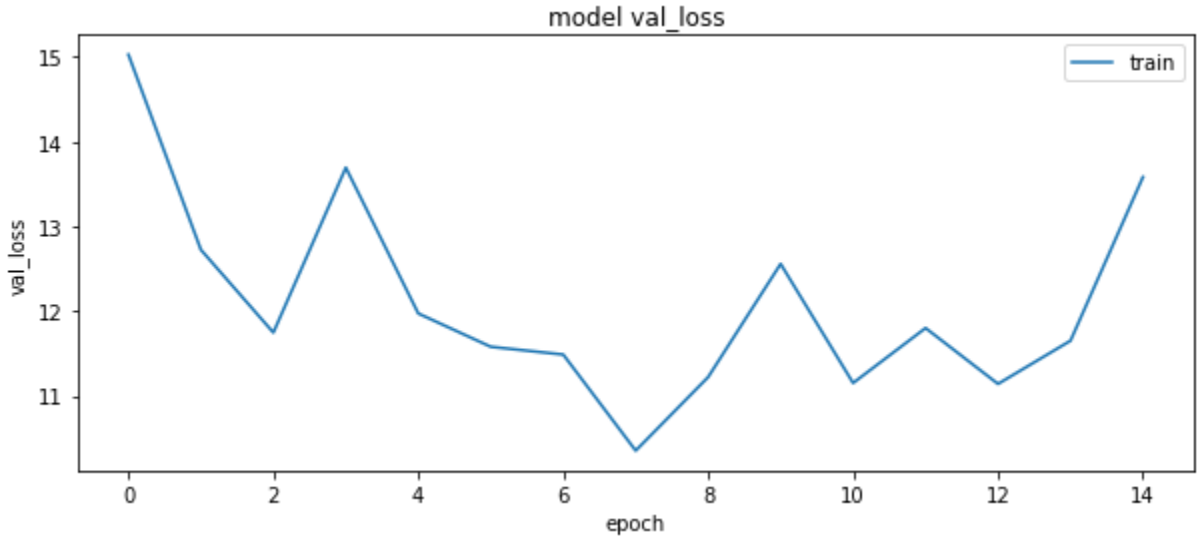
مزیت استفاده از CNN این است که می‌توانند مستقیماً از داده‌های سری زمانی خام یاد بگیرند و به نوبه خود برای مهندسی دستی ویژگی‌های ورودی نیازی به تخصص دامنه ندارند. این مدل می‌تواند نمایش داخلی داده‌های سری زمانی را بیاموزد و به طور ایده‌آل به عملکرد قابل مقایسه با مدل‌های متناسب با نسخه‌ای از مجموعه داده با ویژگی‌های مهندسی شده دست یابد.

معماری شبکه CNN ما به شکل زیر است:



که دارای لایه های Conv 1d , max pooling , flatten ,Dense است.

نتایج این مدل به شکل زیر شد:

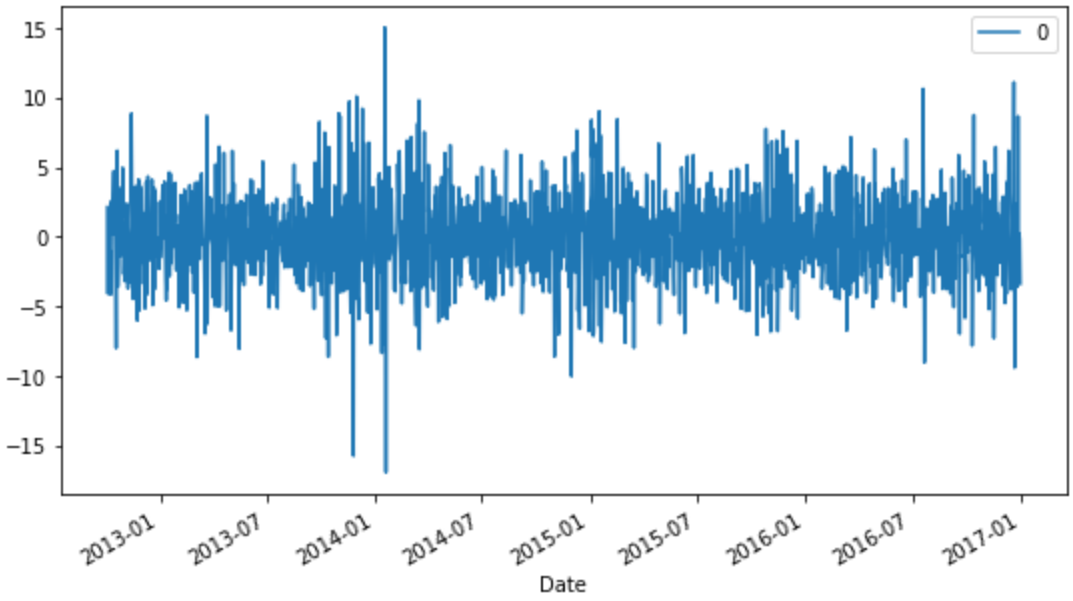


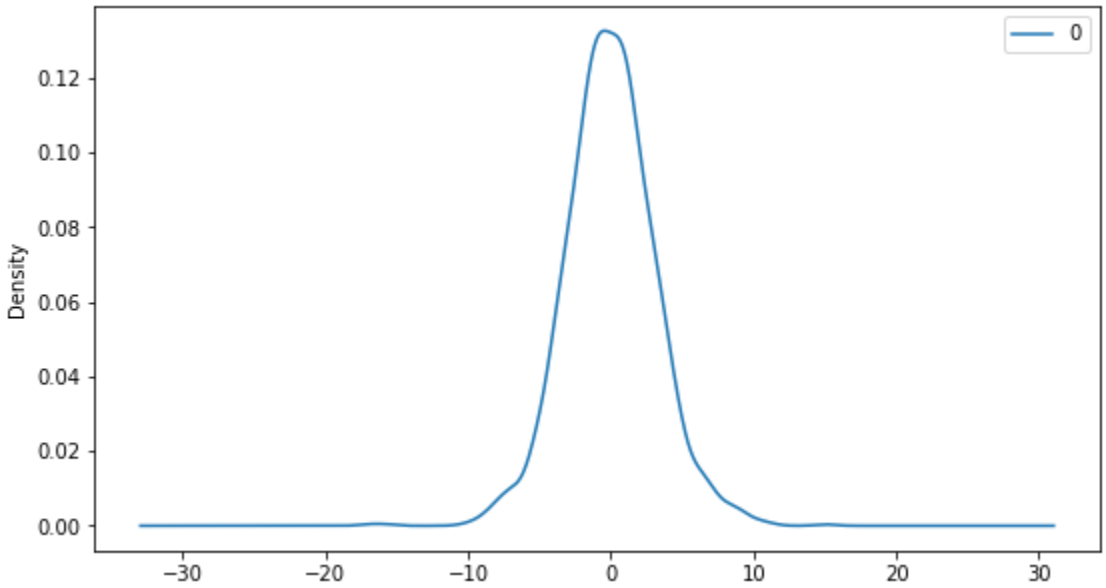
در نهایت loss نهایی روی داده های train 9.22 شد و روی داده های validation 13.58 شد.

3) ARIMA:

یک روش آماری رایج و پرکاربرد برای پیش‌بینی سری‌های زمانی، مدل ARIMA است. این یک کلاس از مدل است که مجموعه ای از ساختارهای زمانی استاندارد مختلف را در داده های سری زمانی به تصویر می کشد. هدف اصلی مدل ARIMA پیش‌بینی (پیش‌بینی مقادیر آتی سری‌های زمانی) است. این مدل به طور کلی به عنوان ARIMA (p، d، q) شناخته می شود، که در آن p، d و q مقادیر عددی غیر منفی هستند و p تعداد sample است و d time step است که نشان می دهد به چند بازه زمانی قبل نیاز دارد و q هم تعداد ویژگی ها است.در این مدل داده های سال 2017 داده های تست و قبل آن داده های آموزشی هستند.

ما در این مدل ARIMA(5,2,0) را بررسی کردیم.

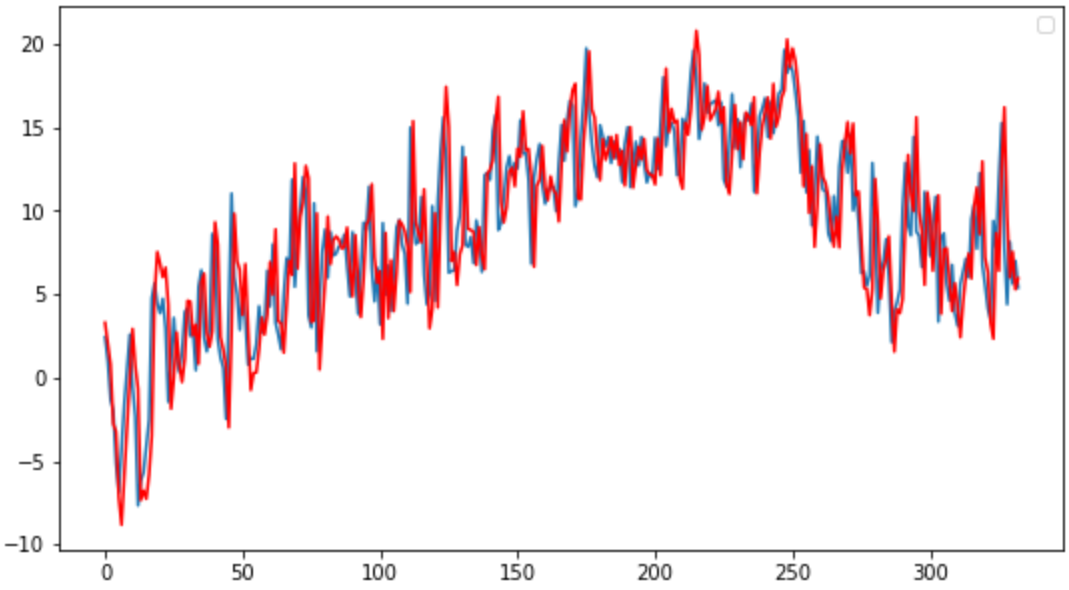




نمودار اول residual های این مدل است در بازه زمانی های مختلف و همینطور نمودار دوم تراکم این residual ها را نشان میدهد.

در نهایت روی داده های تست loss 10.04 را داشتیم.

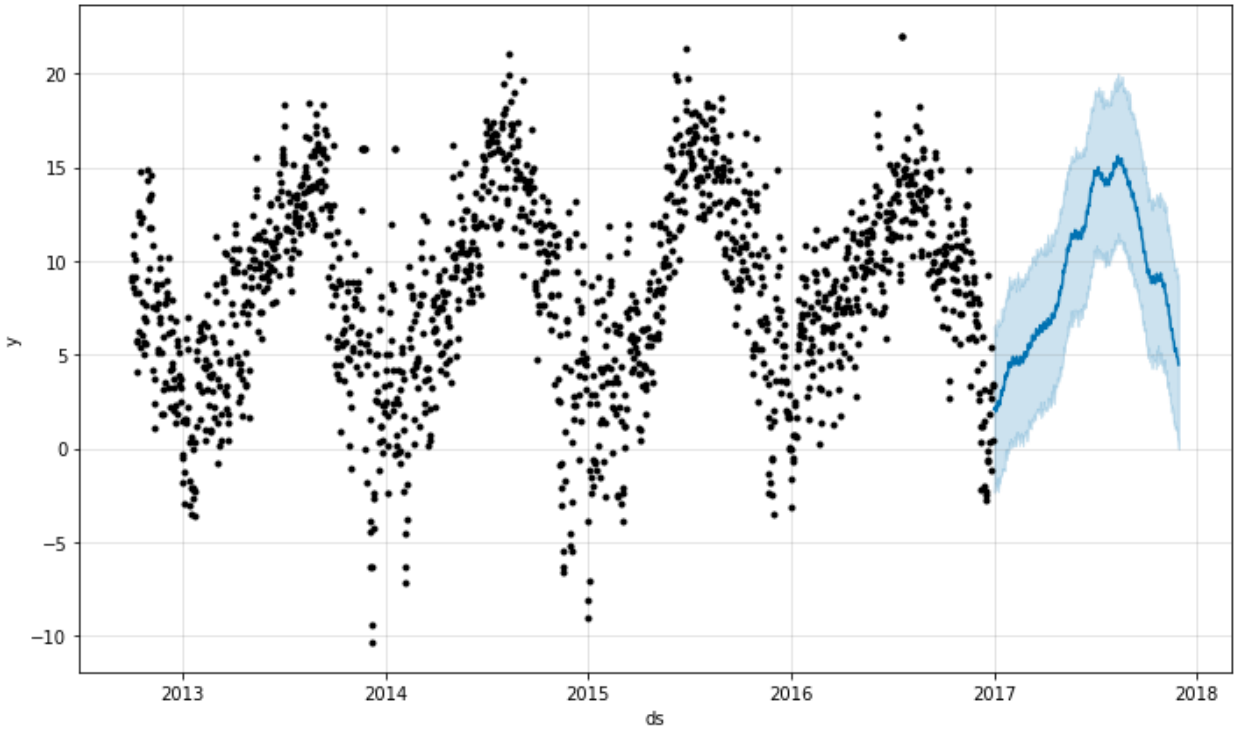
نمودار زیر پیشبینی مدل و داده های واقعی را نشان میدهد:



:Prophet(4

prophet روشی برای پیش‌بینی داده‌های سری زمانی بر اساس یک مدل افزایشی است که در آن روندهای غیرخطی با seasonality سالانه، هفتگی و روزانه به‌علاوه holiday effects مطابقت دارند. این بهترین عملکرد را با سری های زمانی دارد و یکی از نکات مهم prophet این است که نسبت به missing data و تغییرات در روند مقاوم است و معمولاً به خوبی با outliers برخورد می کند. در این مدل هم داده های سال 2017 داده های تست و قبل آن داده های آموزشی هستند.

نمودار زیر داده های واقعی را در کنار داده های پیشبینی شده این مدل نشان می دهد:



در نهایت روی داده های تست loss 9.0 را داشتیم.

در کل با توجه به نتایج بدست آمده بر روی داده های شهر پورتلند بهترین عملکرد را مدل شبکه LSTM داشت و بعد از آن به ترتیب prophet و ARIMA و CNN بودند.
درست است که CNN به تنهایی نتیجه خوبی را نداشت ولی معمولا در کنار مدل های RNN بسیار مورد استفاده قرار می گیرند و حتی برای داده هایی که stationarity ندارند و ناپایدارند(مانند قیمت بیتکوین) مورد استفاده قرار می گیرند و نتایج خوبی دارند چرا که این مدل ها قابلیت درک الگو و pattern را دارند.