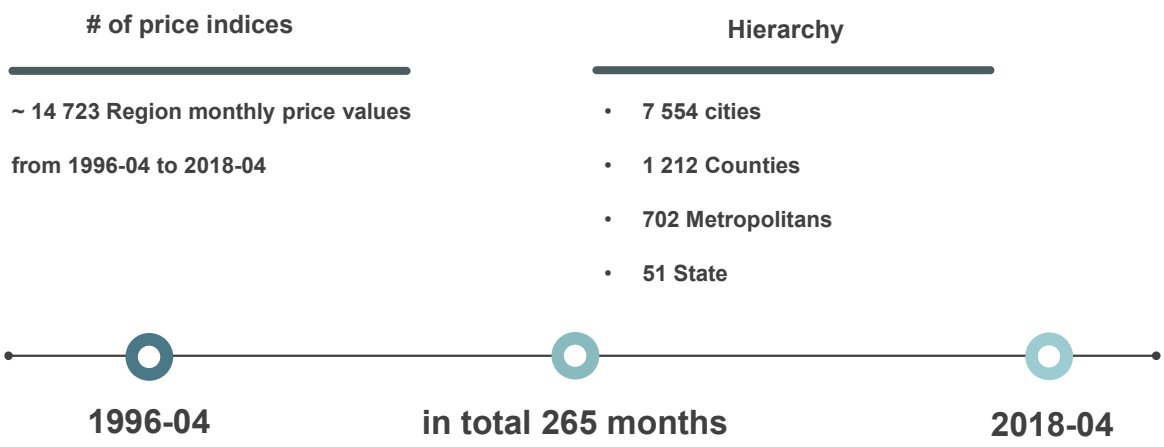
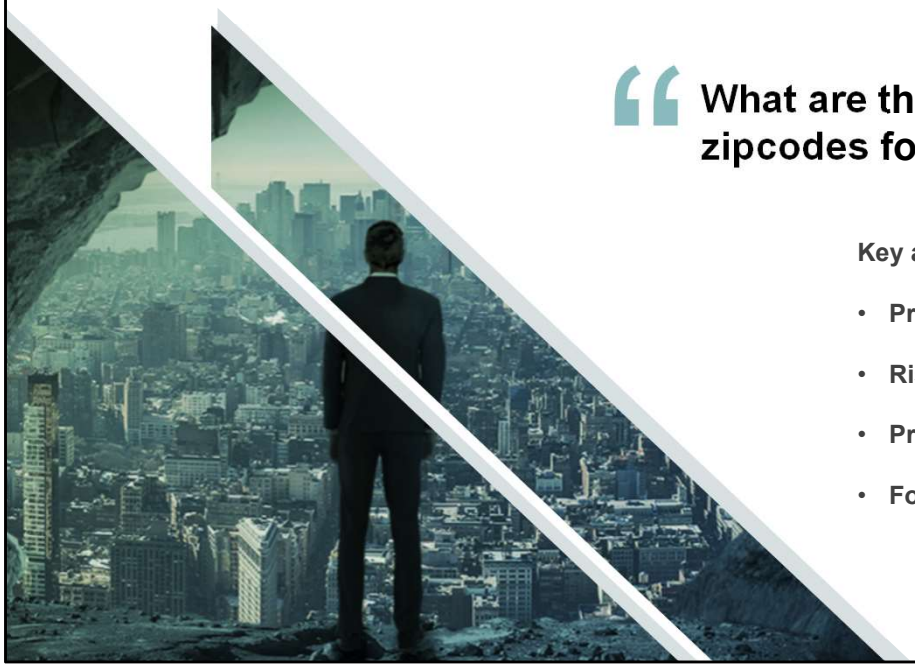
Hi!. Let me present you an Executive Summary of US real estate prices research

Zillow Real Estate Price Dataset



I've used dataset provided by Zillow. It contains price time series per for 14 723 regions. The prices are monthly based and cover a period from April 1996 to April 2018, 265 months in total. The dataset also contains details on geographical location for given regions which are cities, counties, closest Metropolitans and States.

Key Questions



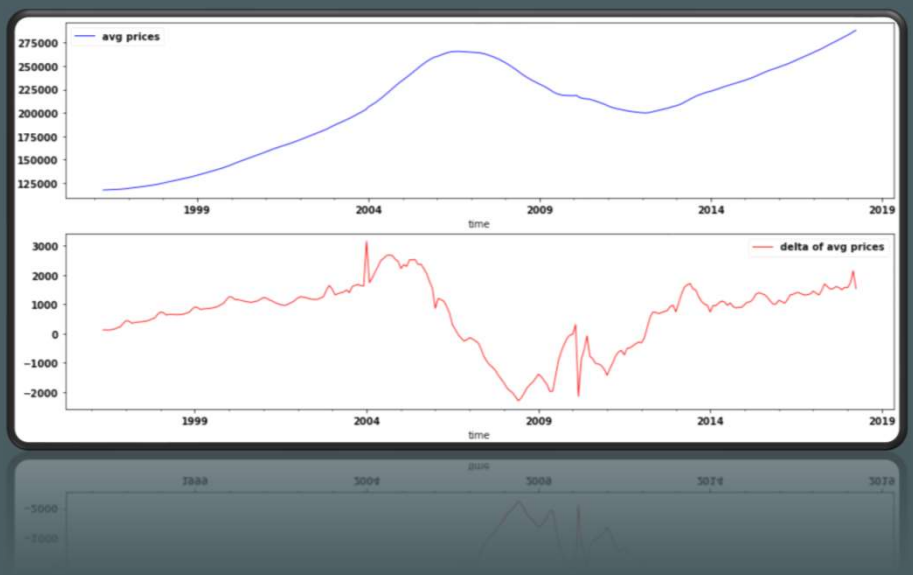
What are the top 5 best zipcodes for us to invest in?

Key aspects

- **Profit metrics**
- **Risk metrics**
- **Profit and Risk trade-off**
- **Forecast period**

Key question for us is to define TOP5 best regions for investing. Before seeking the ways and answer to that, we need to understand what kind of metrics we can use for measure investments' performance. And, since we will use forecasted estimations, we also have to remember about risk measure which leads us to finding trade-off between profit and risk. Depending of model's ability to create robust forecast, we will define recommended forecast period.

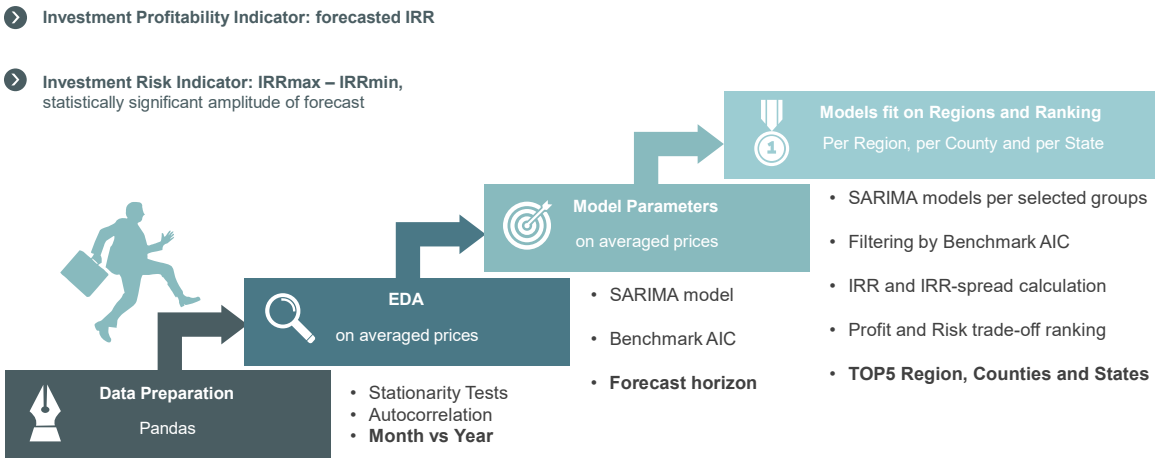
Average prices



- Global cycles (15+ years)
- No significant intra-year fluctuations
- Some intra-year seasonality on price changes

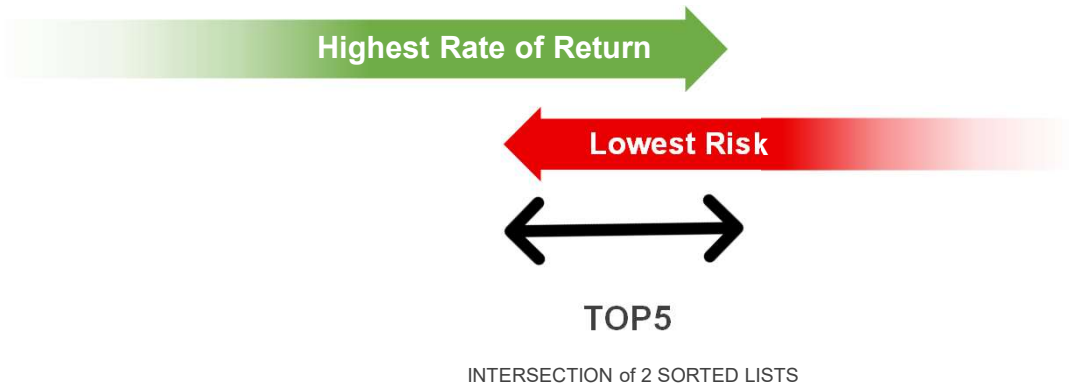
Now, let's see initial prices dynamics. Here you can observe prices and delta prices (monthly changes) based on averaged prices. Firstly, we can notice some global cycles (seasonality) for prices (which could take 15 years periods). I suppose, that it linked to global economic cycles. Secondly, prices are relatively smooth within a year. If we looks to delta prices, we can notice some seasonality within a year. In overall, it is clear, that there is a positive trend for prices, which is expected for such long-term period.

Forecast Methodology and Metrics



A few words about methodology used in the work. I've chosen IRR as estimation of Profitability, which is widely used as KPI for investments for period with length more than 1. Our price forecasts will contain low and high estimations based on statistically significant intervals, and difference between high and low level of IRR is reflect to expected fluctuations of IRR and could be treated as risk measure. In general, I used 4 main steps in the work. After data preparation (technical step), I tested dataset on stationarity and autocorellation in order to understand model specification for next steps. One of important issue here – is to resizing time series from monthly based index to annual based index which not harm outputs, but significantly reduce time for calculation. The third step consists of 1 model runs based on averaged prices. The main outcome here is SARIMA model parameters for next step. In the last step I've run a set of models for each region with fixed params from previous step, calculate IRR and IRR spread and define top5 regions with trade-off mechanism. Additionally, I've calculated same outcomes for Counties and States. More details of methodology is on the APPENDIX A slide.

Profit vs Risk Trade-off



Let me give a brief explanation of trade-off mechanism. Truly speaking, it is quite simple rule: so, we create 2 sorted lists of regions. The first list sorted by IRR in ascending order, the second – by Risk in descending order. Intersection of those lists with same 5 regions – is the answer of TOP5 regions. This is what can be called – balanced choices.

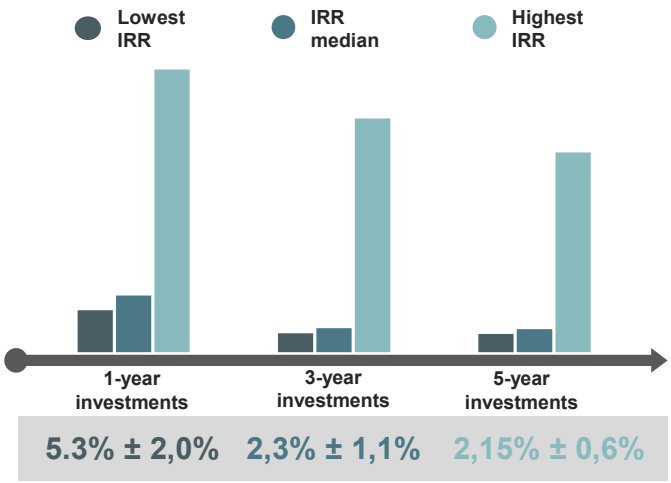
TOP 5 Best Regions

ZIPCODES for 1-year horizon

- 14414 (Avon, NY)
- 14423 (Caledonia, NY)
- 14485 (Lima, NY)
- 19023 (Darby, PA)
- 44102 (Cleveland, OH)

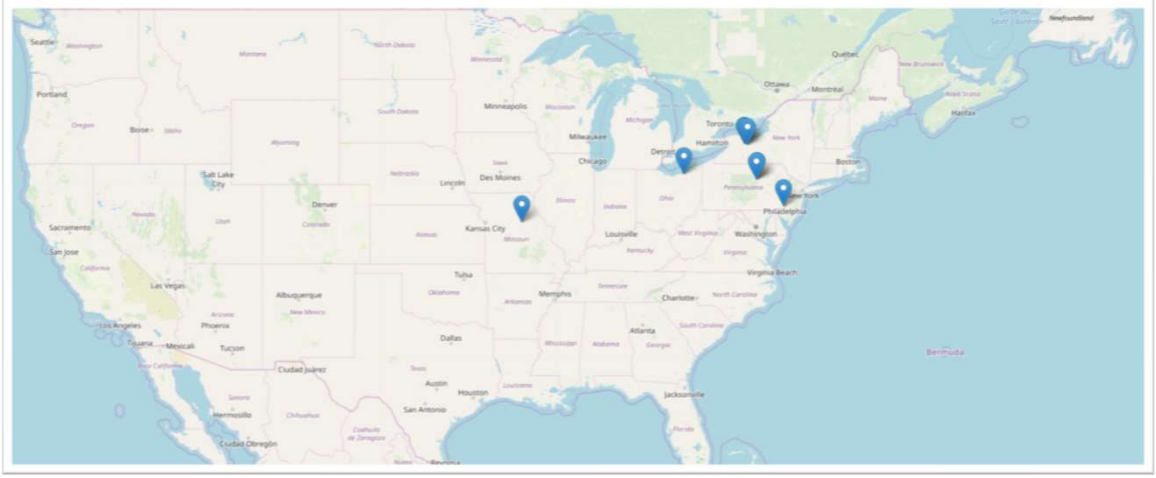
ZIPCODES for 3 and 5-year horizons

- 17701 (Williamsport, PA)
- 17702 (South Williamsport, PA)
- 19023 (Darby, PA)
- 44102 (Cleveland, OH)
- 65255 (Hallsville, MO)



Here you can see results of modelling and ranking. So, we have 2 sets of best balanced regions: for 1-year investments and 3-5 year investments. On the right side of presentation, IRR estimations are presented, where Lowest IRR and Highest IRR were taken from forecast bounds. For 1-year payback investments in average 5,3% IRR could be obtained from top5 regions. Average amplitude is 2% (half of spread). Some important issue here: profit measure is decreased with increasing forecast length, which happens because decreasing forecast power with increased length.

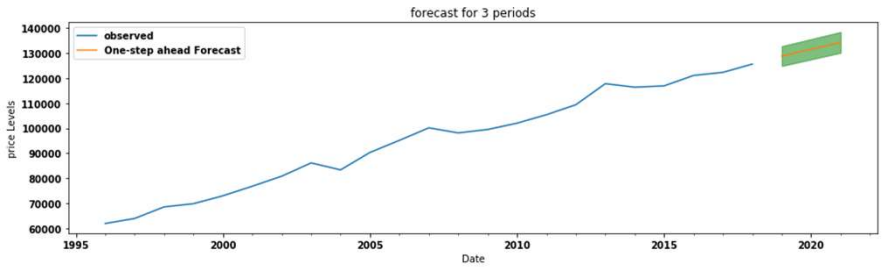
TOP 5 Best Regions: map



TOP ZIPCODES: 44102, 17701, 17702, 19023, 65255, 14414, 14423, 14485

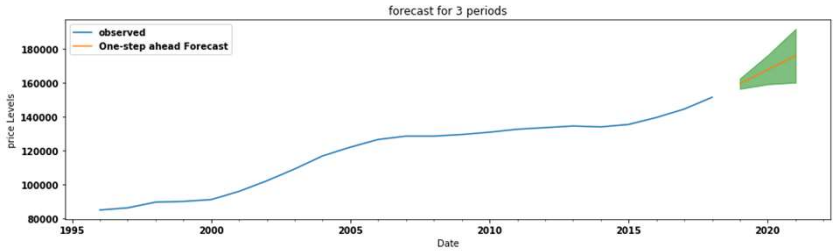
Here you can see locations of TOP5 best zipcodes in the MAP. The overall number of regions is higher than 5 because I put here TOP5 for 1-year and TOP5 for 3-year periods.

3-year Forecast Charts



<= 17701 (Williamsport, PA)

14414 (Avon, NY) =>



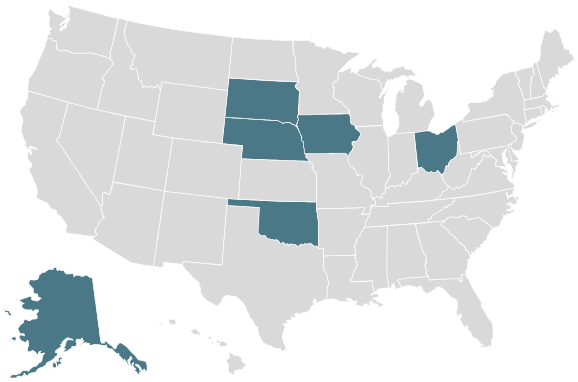
Here, I've put 2 examples (region 17701 and 14414) of how forecasts looks like. You can see, that shape of confidence intervals could be different for different regions and it depends on individual data for a region.

TOP5 Best Counties and States

COUNTIES		STATES	
1-year	3-year	1-year	3-year
Allegheny	Chemung	Iowa	Iowa
Erie	Erie	Nebraska	Nebraska
Maury	Maury	Ohio	Ohio
Pontonoc	Pontonoc	Oklahoma	Alaska
West Baton Rouge	Dubuque	South Dakota	South Dakota
4.4% ± 2.4%	4.1% ± 2.3%t	4.0% ± 2.3%	3.8% ± 3.3%

In this table you can figure out TOP5 sets for Counties and States. The main idea is that increasing location scale leads to decrease estimated IRR, which is logically correct, because upscaling datasets decrease mean values of prices and IRR.

TOP5 Best States



TOP 5 STATES:

- Alaska
- Oklahoma
- Ohio
- Iowa
- South Dakota
- Nebraska

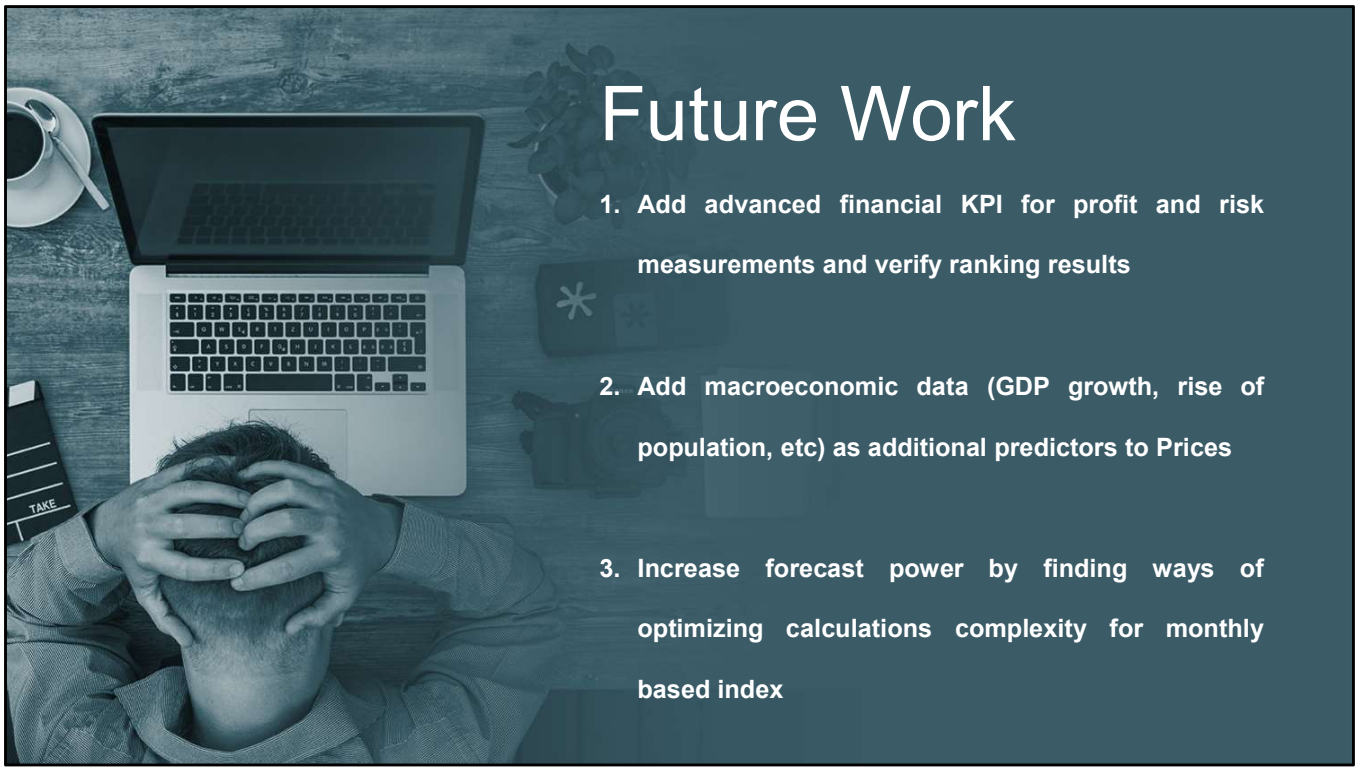
In total 6 states as union of 1-year and 3-year lists

In this slide you can see the location of best states in the US map.

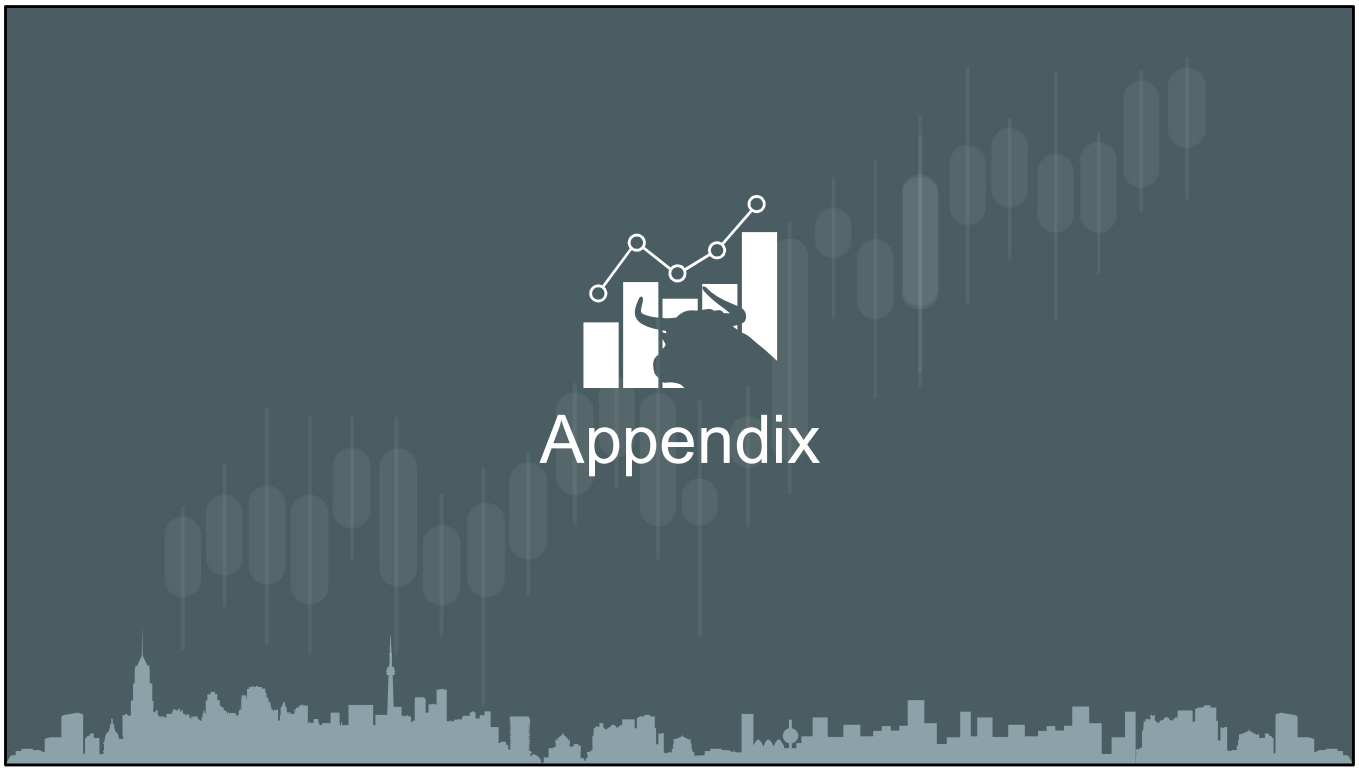
Business Recommendations

- Prices on the given dataset **could be forecasted** with **high and moderate** level of accuracy
- Recommended forecast **length: 5 years or less**
- Top5 **zipcodes** for investments:
  - for **1-year** maturity: **14414, 14423, 14485, 19023, 44102;**
  - for **3-5-year** maturity: **17701, 17702, 19023, 44102, 65255**
- Median **IRR** for TOP5 **zipcodes**: **5,3% (1-year), 2,3% (3-year), 2,1% (5-year)**
- Top5 counties for investments (3-year): **Erie, Keweenaw, Maury, Pontonoc, Dubuque**
- Top5 states for investments (3-year): **Alaska, Iowa, Nebraska, Ohio, South Dakota**

So, let me state main findings and recommendations. The provided dataset is gives us the opportunity to create predict models for prices. I can say, that obtained results have high and moderate level of prediction quality and forecasts should be limited by 5 years length. Based on IRR metrics we can define top5 regions as it shown in the conclusion slide with IRR varied from 5,3% to 2,1%. Additionally, you can figure out top5 counties and top5 states for investing.



Of course, the given executive summary shows a snapshot of current discovery works. And there are many ways to expand and impove forecast quality. In this slide I've 3 general point to get that. The first is about KPI. Applied indicator is the thing highly depends on business aspects. For example, if we concentrate on 1-year investments, there are many other indicators for profitability measure. The second point is about adding some local (per State) or global (per Country) macroeconomic data to the dataset as possible price predictors. Remember 15-years cycles in price chart. So, such data could be the answer for that. And the third point. In current state, running the code, but on monthly based data – is extremely rise calculation time. On the other hand, forecast accuracy could be increased. So the task here – is to find optimal way with low cost.
Thank you for attention!


Appendix

Appendix A. Detailed Calculation Steps

A. EDA

1. Check stationarity (ADF test)
2. Check ACF and PACF
3. Check monthly ts vs rolling ts with 12month
4. Choose resize option (**YEAR**) and range for SARIMA params (**0,1**) – as pqd and **s=1**

B. Benchmark SARIMA

1. Prepare TS with avg-prices
2. Find optimal SARIMA-params **(0,1,1)(0,1,1,1)**
3. Calculate benchmark AIC (**376**)

C. Prepare data for model

1. For original question: row ~ Region
2. Additional 1: row ~ County, prices – averaged
3. Additional 2: row ~ State, prices – averaged
4. Time period for (1)-(3) - **Year**

D. Models Fit

1. Run all TS with 3 DF with fixed SARIMA params (from B)
2. Filter results with Benchmar AIC (see in B) and collect to DF
3. Calculate IRR and IRR-spread for each filtered row and add to DF

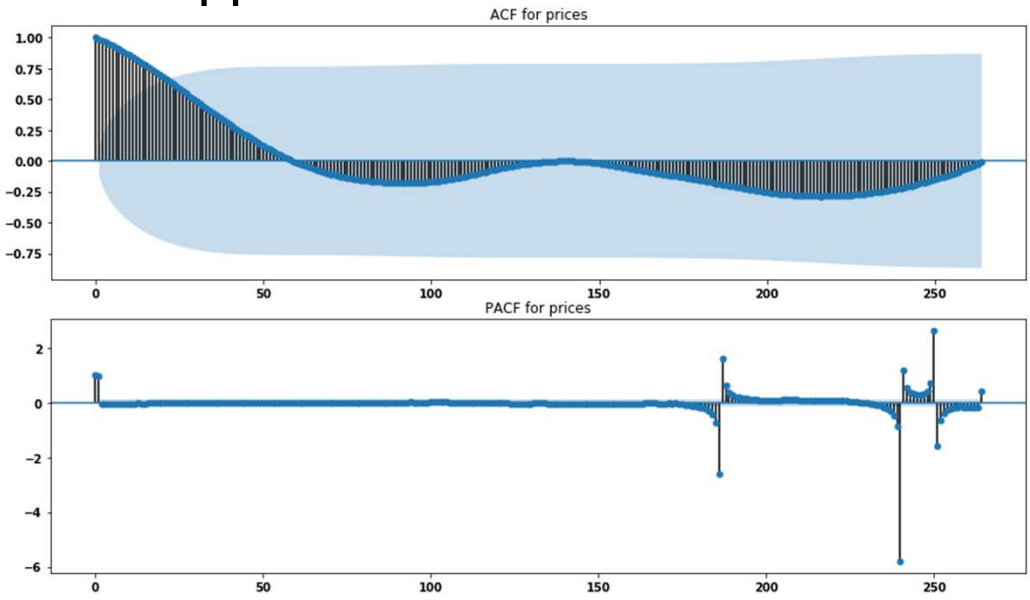
E. Ranking

1. Rank all rows (Regions, Counties and States) in 3 DF by IRR and IRR_spread
2. Collect top best (5) results and put to resulting DF

E. Advanced tests

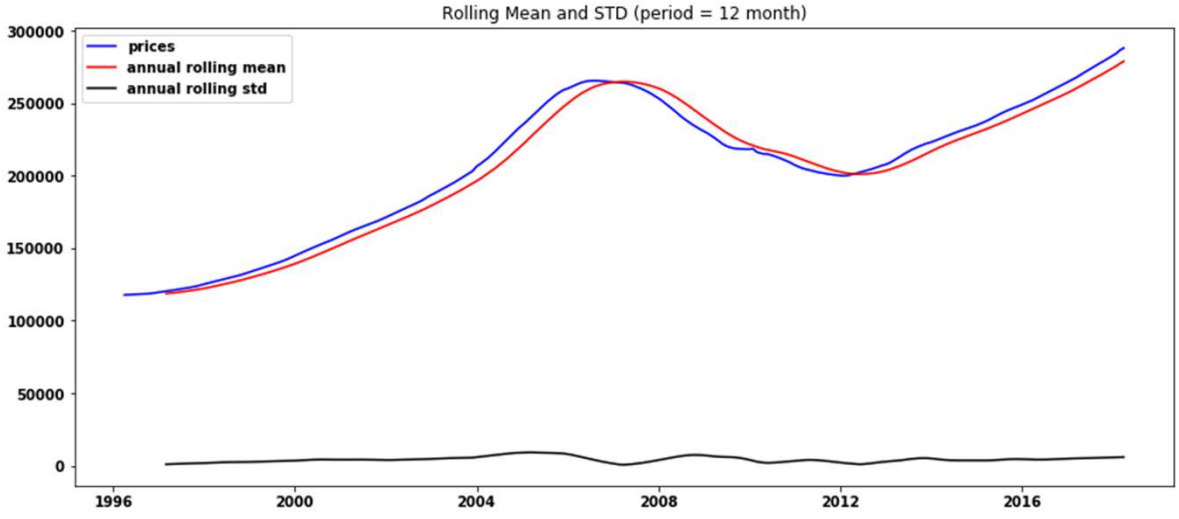
1. Calculate and compare with forecast charts from E a set of models with different optimal SARIMA params for each row
2. Run and compare Monthly Based model and forecast with results from prev.steps (annualy based)

Appendix B. ACF and PACF



The charts above calculated on average prices

Appendix C. Month vs Year

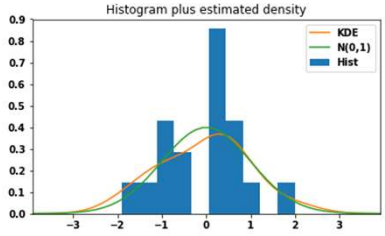


The charts above calculated on average prices

Appendix D. Benchmark model and forecast

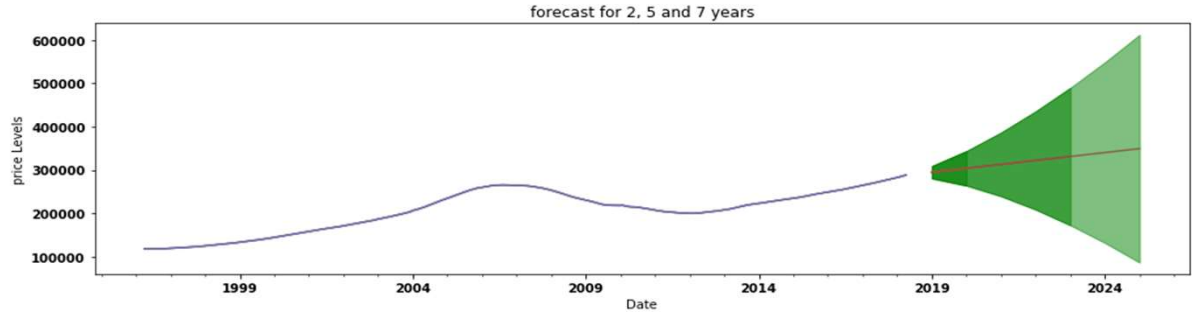
Best Model Params:

- Data: **average price p.a.**
- pdq: **(0,1,1)**
- PDQs: **(0,1,1,1)**
- AIC: **376**



Residual params and Validation:

- P-values: **less than 1%**
- Skew: **-0.03**
- Kurtosis: **2.56**



The chart above calculated on average prices