Predicting Housing Prices

By Dustin Stewart

The Problem

How can we make good real estate investments if we can't accurately predict the selling price of a property?

Problem

How can we make good real estate investments if we can't accurately predict the selling price of a property?

Objective

Create a regression model that can accurately predict the selling price of a home

Dustin
Stewart

About Me

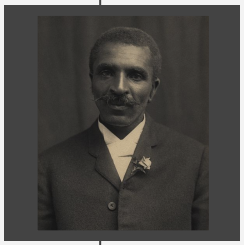
Data scientist at General Assembly and have a background in digital marketing as well as finance.

Started doing data analysis back in 2017 while studying at the University of Southern California

Table of Contents

01 The Data

Data Background



70 Columns

2006-2010

2051 Rows

Ames, Iowa

Population: 55,000

Home to Iowa State University, founded in 1858

The Data

Details of over 2000 houses sold between 2006 and 2010 in Ames

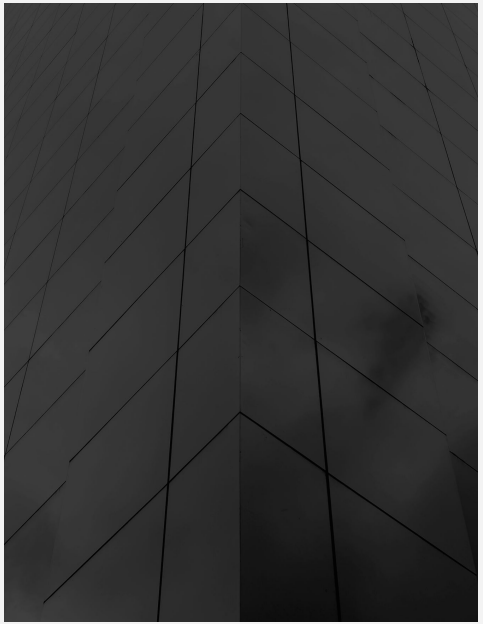
The data starts during the last years of the housing bubble and continues for two years after it burst

Data Cleaning

1. Replace null values while preserving data integrity

2. Reformat categorical and boolean variables so they can be used in model as effectively as possible

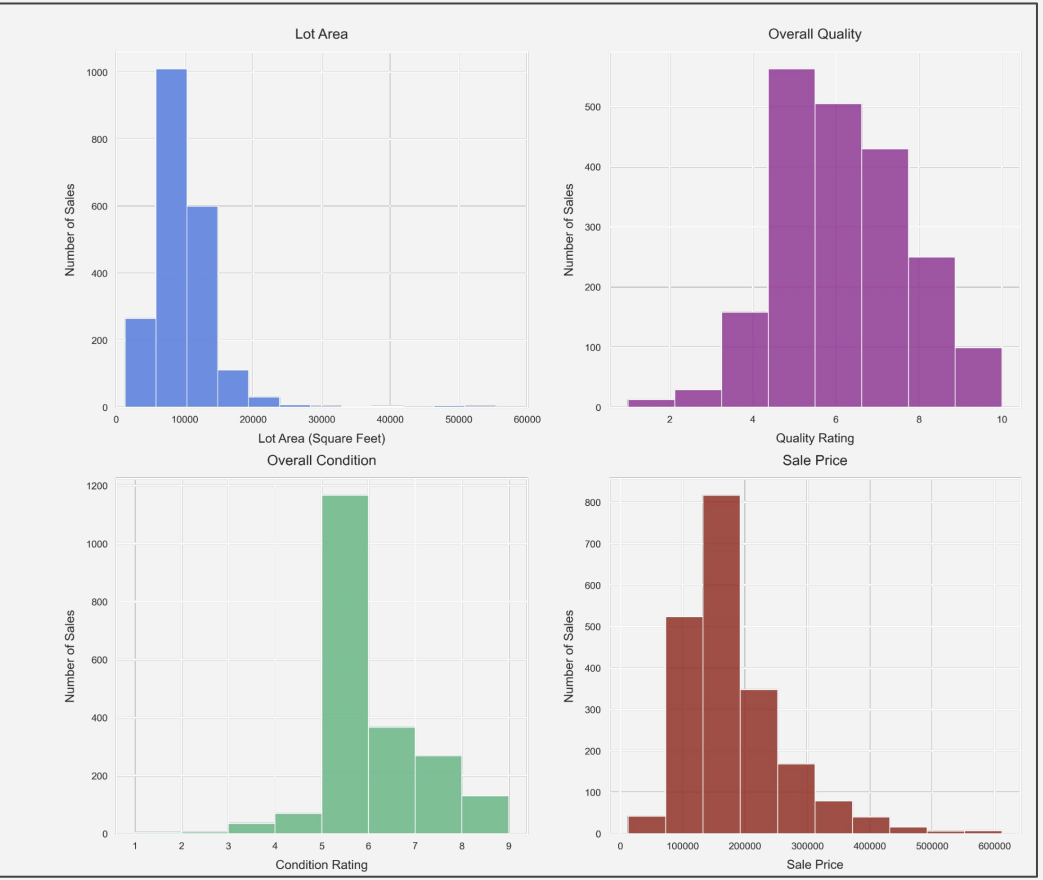
3. Clean further as exploratory analysis demands



Exploring the Data

02

Methodology

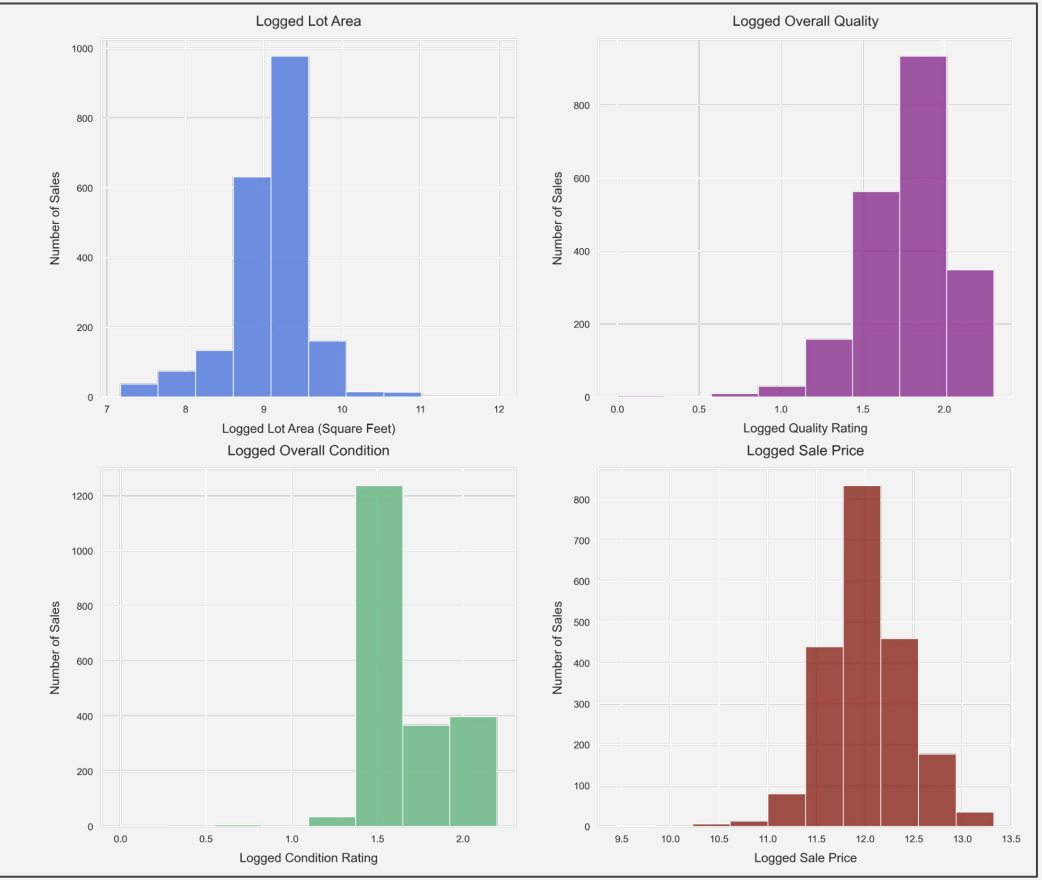


Analyzing Distributions

We see that certain distributions appear to be skewed right

A common way to make the distribution more normal is to use a logarithmic transformation

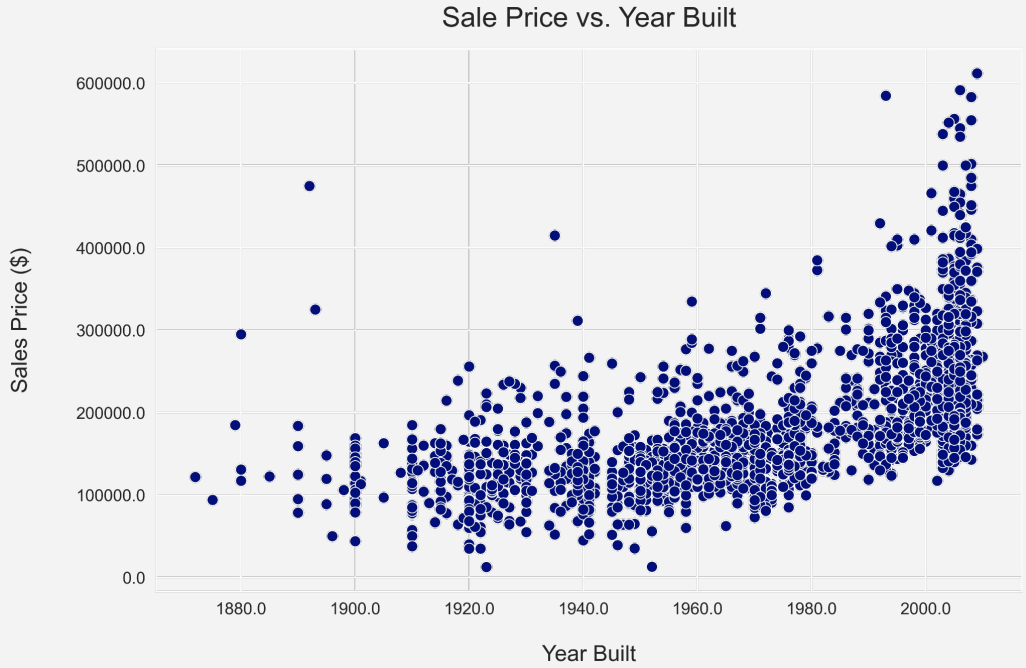
Methodology



Sales Price and Lot Area are definitely more normally distributed when logged

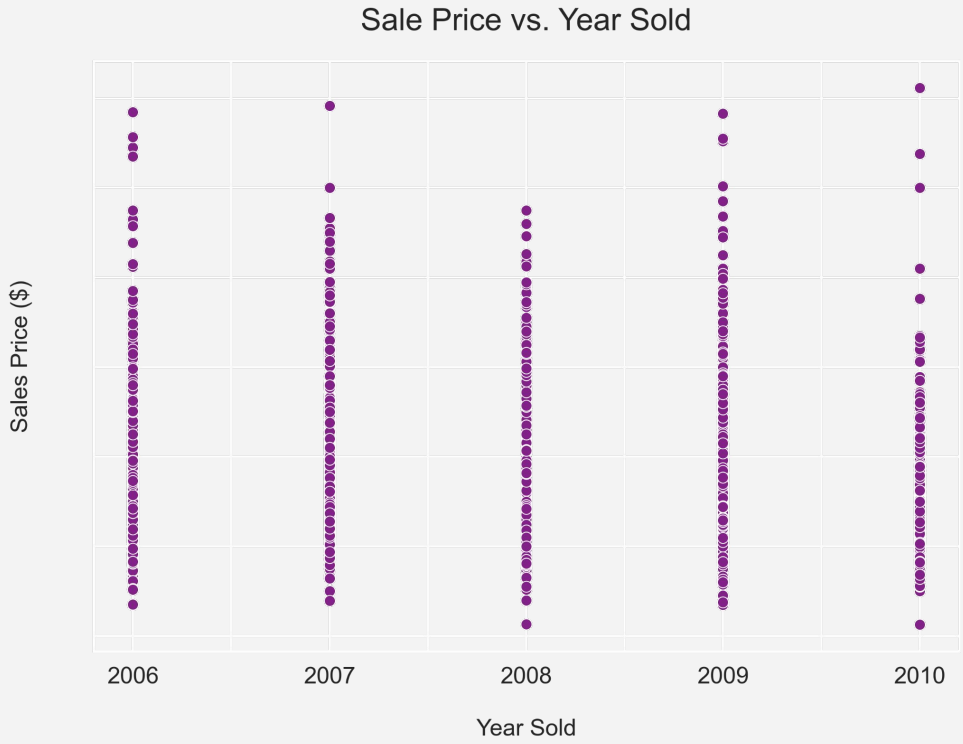
Scatter Plot Analysis

Why are there so many outliers towards the bottom of this distribution?

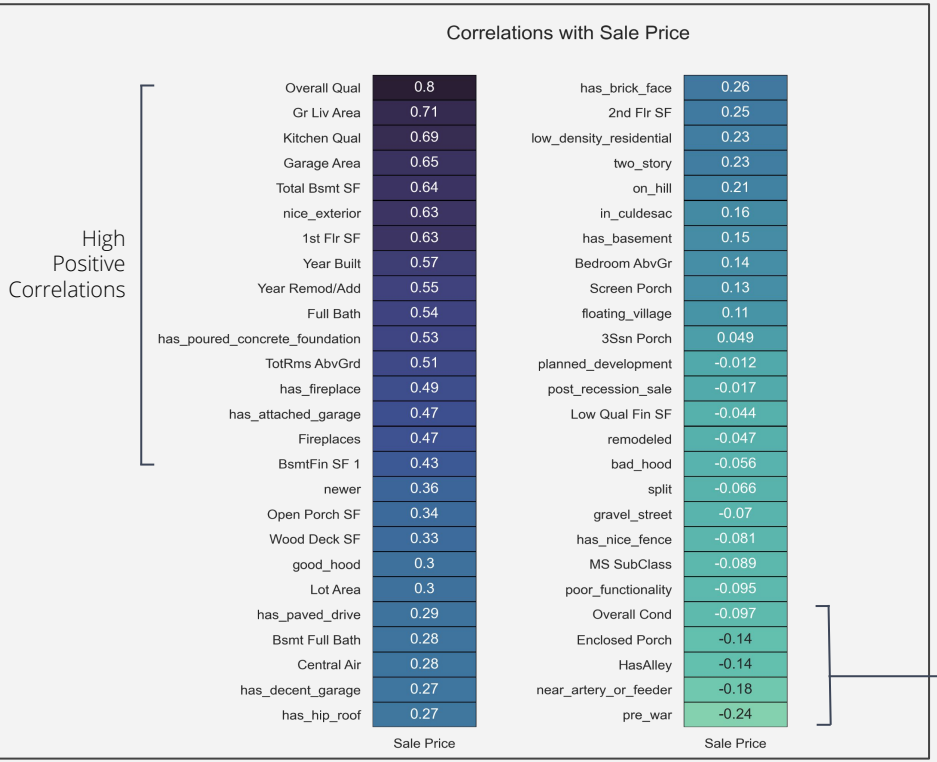


Scatter Plot Analysis

How did the bursting of the housing bubble affect Ames?



Methodology



Feature Selection

Looking at correlations between different variables and sales to help choose which to use for the baseline model.

Continually return to these correlations to figure out why a variable's weight in the model doesn't reflect its level of correlation with sales price.

03 Creating a Model

Model Selection

	Linear Regression	Lasso	Ridge
			
R2 Score =	**0.904**	**0.905**	**0.904**
	This is the basic regression model	Lasso adds a penalty that increases as the weight of the variable increases	Ridge also adds a penalty but will never make a coefficient absolute zero like Lasso does

Baseline Score

We calculate a baseline error to compare our model to by scoring the mean sale price from our data

Error = **$79,282**

First Model Score

We run all of our numeric variables through the model:

Training Error = **$20,801**
R2 Score = **0.91**

Wow what an improvement!

Testing Error = **$95,373**
R2 Score = **0.83**

Refining the Model

04

Methodology

	Features	Coefficient
0	Overall Qual	0.117359
1	Year Built	0.079079
2	BsmtFin SF 1	0.035522
3	1st Flr SF	0.105314
4	2nd Flr SF	0.085343
5	Kitchen Qual	0.031546
6	remodeled	0.002605
7	post_recession_sale	-0.002916
8	HasAlley	-0.003548
9	has_brick_face	0.003637
10	has_decent_garage	0.021931
11	floating_village	0.011741
12	low_density_residential	0.008183
13	gravel_street	-0.008023
14	on_hill	0.004157
15	in_culdesac	0.001034
16	near_artery_or_feeder	-0.016018
17	has_hip_roof	0.006775
18	nice_exterior	0.000684
19	has_poured_concrete_foundation	0.010510
20	poor_functionality	-0.020989
21	has_fireplace	0.023629
22	has_porch_or_deck	0.006814
23	log_overall_cond	0.055932
24	log_lot_area	0.053222
25	Neighborhood_Blueste	-0.000000
26	Neighborhood_BrDale	-0.003984
27	Neighborhood_BrkSide	-0.000000
28	Neighborhood_ClearCr	0.006302
29	Neighborhood_CollgCr	-0.000000
30	Neighborhood_Crawfor	0.015979
31	Neighborhood_Edwards	-0.019176
32	Neighborhood_Gilbert	-0.003449
33	Neighborhood_Greens	0.000000
34	Neighborhood_IDOTRR	-0.021182
35	Neighborhood_MeadowV	-0.003553
36	Neighborhood_Mitchel	-0.000000
37	Neighborhood_NAmes	0.000000
38	Neighborhood_NPkVill	-0.000000
39	Neighborhood_NWAmes	-0.003826
40	Neighborhood_NoRidge	0.006992
41	Neighborhood_NridgHt	0.018654
42	Neighborhood_OldTown	-0.015234
43	Neighborhood_SWISU	0.000000
44	Neighborhood_Sawyer	-0.000949
45	Neighborhood_SawyerW	-0.006022
46	Neighborhood_Somerst	0.007245
47	Neighborhood_StoneBr	0.013030
48	Neighborhood_Timber	0.000853
49	Neighborhood_Veenker	0.000000

Recursive Feature Selection & Creation

By:

A. Calculating the significance of each variable's effect on the model
B. Comparing their effect on the model to correlations
C. Using a bit of trial and error

We can find the variables that best predict sale price

Final Model Criteria

Independent Variables

- Overall Quality
- Overall Condition (log)
- Lot Area (log)
- Year Built
- Finished Basement SF
- 1st Floor SF
- 2nd Floor SF
- Kitchen Quality
- Remodeled
- Sold Post Recession
- Has an Alley
- Has Brick Face
- Has a Decent Garage
- On a Gravel Street
- On a Hill
- In a Cul-de-Sac
- Near an Artery or Feeder Street
- Has a 'Hip' Roof
- Has a Nice Exterior
- Has a Poured Concrete Foundation
- Has Poor Functionality
- Has a Fireplace
- Has a Porch or Deck
- In the Floating Village
- In a Low Density Residential Area
- Neighborhoods

05 Model Insights

R2 Scores

Training = **0.89**

Test = **0.90**

RMSE Scores

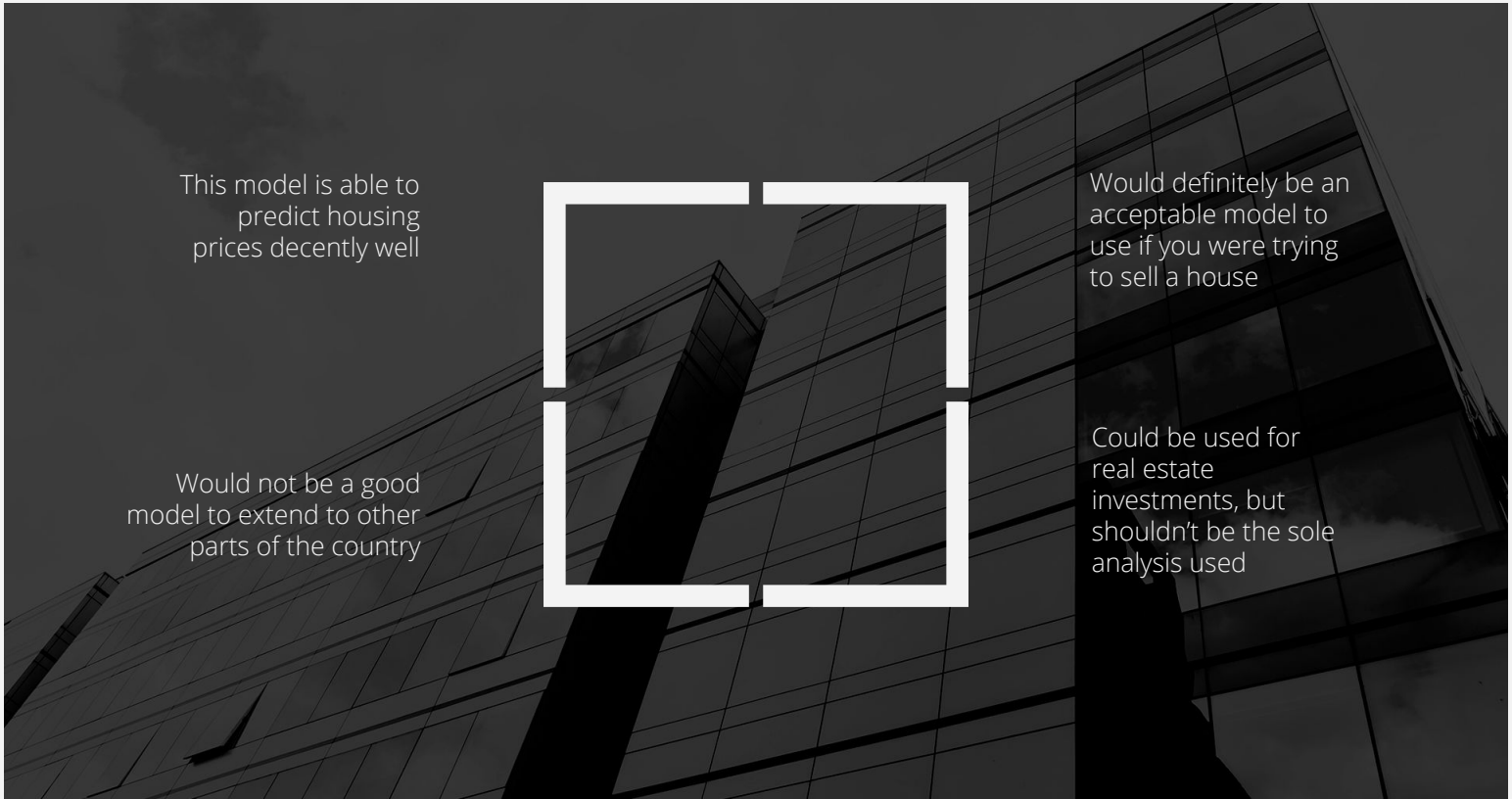
Training Error = **$36,301**

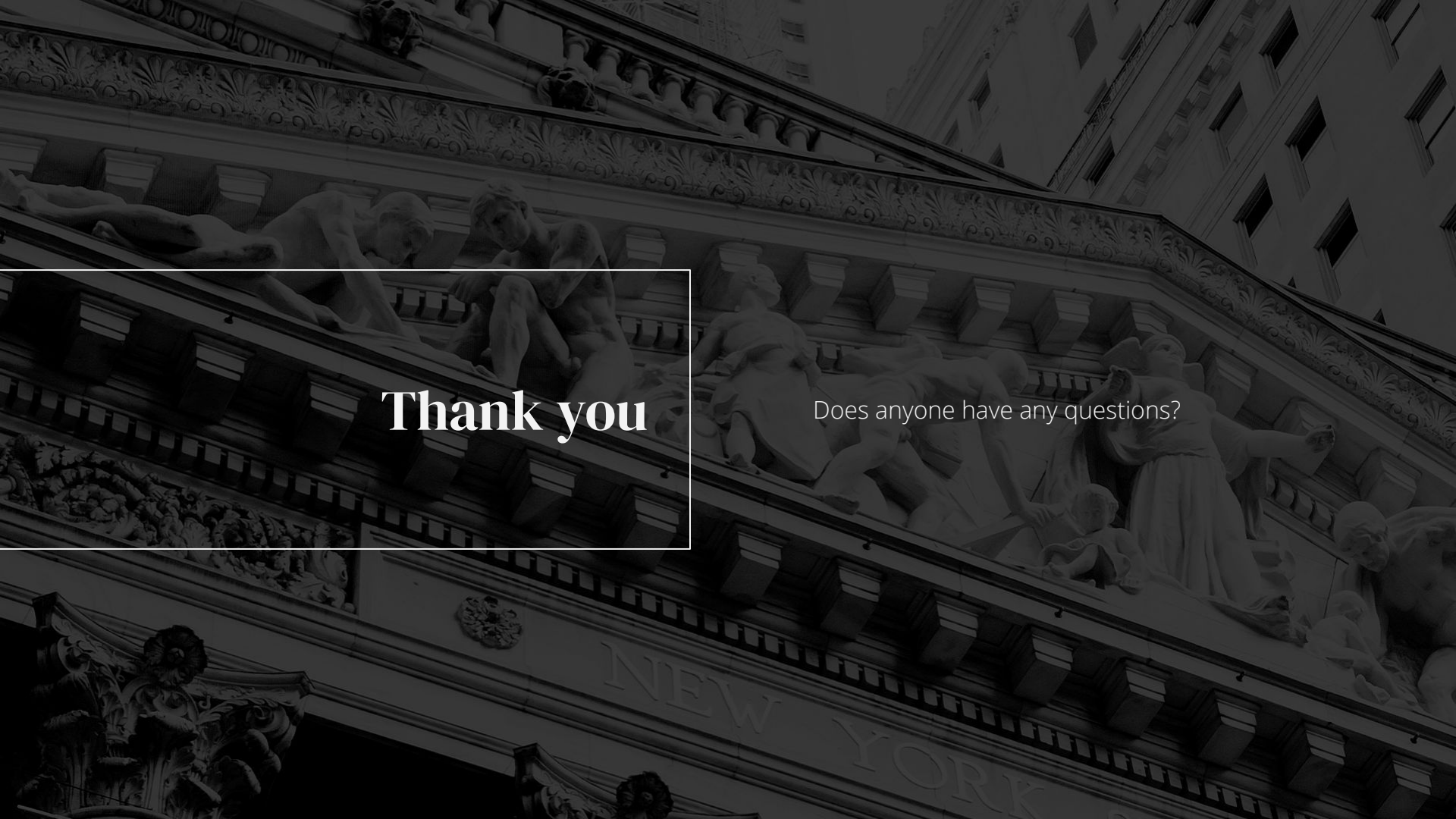
Testing Error = **$20,691**

06

Conclusions

Conclusions





Thank you

Does anyone have any questions?


slidesgo