Work Breakdown

|  |  |
| --- | --- |
| **Item** |  |
| **Phase 1: Data Preparation** |  |
| **Primary Datasets** |  |
| Philadelphia Property Sales (cleaned)   * Documentation/narrative of cleaning decisions * Summary tables of dimensions before/after cleaning |  |
| **Secondary Datasets** |  |
| Census   * Documentation/narrative of cleaning decisions * Joined to Sales Data * Summary tables of dimensions before/after cleaning |  |
| OpenDataPhilly   * Documentation/narrative of cleaning decisions * Joined to Sales Data * Summary tables of dimensions before/after cleaning |  |
| **Phase 2: Exploratory Data Analysis** |  |
| **5 Professional Visualizations** |  |
| Distribution of sale prices (histogram) |  |
| Geographic distribution (map) |  |
| Price vs. structural features (scatter plots) |  |
| Price vs. spatial features (scatter plots) |  |
| One creative visualization |  |
| **Choose 2-3 Visualizations for Presentation** |  |
| Visualization 1   * Caption/description |  |
| Visualization 2   * Caption/description |  |
| **Phase 3: Feature Engineering** |  |
| **Features** |  |
| Buffer-based features (Parks & rec sites within 0.25 mi, Crime counts within 0.5 mi – log transformed, Schools within 0.5 mi)   * Creation code & justification for feature | Kavana |
| k-Nearest Neighbor feature (Distance to nearest transit stop, Distance to nearest hospital)   * Creation code & justification for feature | Kavana |
| Census variables (population, med hh inc, med age)   * Creation code & justification for feature |  |
| Interaction terms (theoretically motivated combinations of variables) ()   * Creation code & justification for feature |  |
| **Summary table of features created** |  |
| **Phase 4: Model Building** |  |
| **Model Creation** |  |
| Model 1 - Structural Features ONLY   * Model code, stargazer table output, coefficient interpretation text |  |
| Model 2 – Add Census Variables   * Model code, stargazer table output, coefficient interpretation text |  |
| Model 3 – Add Spatial Features   * Model code, stargazer table output, coefficient interpretation text |  |
| Model 4 – Interactions and fixed effects   * Model code, stargazer table output, coefficient interpretation text |  |
| **Model Comparison Table**   * RMSE * R2 |  |
| **Phase 5: Model Validation** |  |
| 10-fold cross-validation comparison between all 4 models   * Code * Detailed results table * Predicted vs. actual scatter plot * Discussion of which features matter most to the models |  |
| CV results table |  |
| One compelling visual |  |
| **Phase 6: Model Diagnostics** |  |
| 3 Diagnostic Plots   * Residual plot (linearity, homoscedasticity) * Q-Q plot (normality) * Cook’s distance (influential observations) * Interpretations of each plot * If we address violations of these tests, text description of how we did |  |
| **Phase 7: Conclusions & Recommendations** |  |
| **Questions to Answer**   * Final model accuracy * Which features matter most for Philadelphia’s home prices? * Which neighborhoods are hardest to predict? * Equity concerns? * Limitations? |  |
| **Appendix** |  |
| 2-3 paragraph detailed discussion |  |
| **Presentation** |  |
| 1-2 slides with clear, concise answers |  |

Technical Appendix -

Quarto Presentation Slides (10-15) – Nina

Work Schedule

* Wednesday – CR (afternoon), HSH (afternoon), KR (afternoon), RD (afternoon)
  + **Complete Phase 2 & 3**
    - Clean & analyze features for possible implementation methods
* Thursday – HSH (morning), CR (morning), NC (morning), SK (not python)
* Friday – CR (~morning), SK (anytime)
  + Complete Phase 4 by EOD
* Saturday – SK (anytime), KR (anytime), HSH (midday), RD (anytime), NC (anytime)
  + Complete Phase 5 & 6
* Sunday – SK (quick meeting), KR (anytime)
  + Complete Phase 7
  + Practice presentation
* Monday
  + Presentation Day!