# Technical test in Data Science

Hello candidate. Congratulations on reaching this stage!

We look forward to seeing more of your ability to analyze, transform, and process data. Any questions that arise during the process, please feel free to contact [recrutamento@datasprints.com](mailto:recrutamento@datasprints.com). Shall we?

## Goals

The purpose of this test is to assess your proficiency in **Data Analysis**. For that, you will need to offer insights and infer results by sampling the data provided, using either graphs or comments, and presenting them through [storytelling](https://www.columnfivemedia.com/data-storytelling-brands-data-visualization) techniques.

The **minimum skills** for our job opportunity are:

* Basic Python programming;
* Basic R programming;
* Statistics, Machine Learning, Optimization and Data Mining knowledge;

Also, **differentiation skills**:

* Cloud Computing experience;
* Linux experience;

Therefore, focus on showing that you **meet our criteria** by answering the [minimal requirements](#_dyfc3cx7ue8e) using a mix between *Python/R*. Leave to cover the differentials in the points raised in [bonus items](#_t7e1gpy2296s).

## Deadline

You will have **7(seven) days** to submit this test **from the date this email was sent**. We consider that you have at least 1(one) weekend of effort to complete it. If you want more time, please contact us to set a deadline that best fits us all. :)

## About the data

The data was collected by a Data Sprints crawler from multiple portals and is composed of real estate ads. This customer is an ad scraping company and needs to add intelligence to unstructured data obtained through numerous data sources (portals).

We provide you with data gathered by our crawler in the following file:

|  |  |
| --- | --- |
| **Dataset** | **Description** |
| [realties.json](https://drive.google.com/file/d/1-AkngJMOej-lyxLaDTnryYPowJbK4E3P/view?usp=sharing) (tar.gz format) | Primary sample gathered by our crawler |

The dataset has issues that should be detected or addressed. To help you, we offer a few items that might give you guidance for your analysis:

* How to identify duplicate data, so the analysis is not skewed?
  + Who duplicates the data?
    - Popular Advertiser?
    - Third company?
    - The company crawler?
* What do these duplicate ads contain?
* When are ads duplicated?
* What happens to duplication, and where is it stored?
* Why are they duplicated?
* What types of interesting outliers can be identified?
  + What rule should we use to remove or maintain outliers?
* The field **lista\_fotos** has links from ad's photos, and some links still work. You can use these photos to try, with visual computing tools, to recognize similar visual patterns or pictures of repeated ads

## Data Structure

The database has the following files:

* **realties.json:** Main sample with data content.

The file **realties.json** has the following fields:

* **\_index**: index
* **\_type**: type (Categories: Real State)
* **\_id**: Individual Identification
* **\_score**: Crawler metrics / elastic net
* **\_source**: data source
  + **valor**: Property value
  + **descricao**: Advertisement Description(in portuguese)
  + **area\_util**: Useful floor area (in m²)
  + **quarto**: Number of bedrooms
  + **uf**: State
  + **latitude**: Property Latitude
  + **longitude**: Property Longitude
  + **tipo\_imovel**: Property Type ( apartment, house, living room, land, etc)
  + **lista\_fotos**: Photo URL list
  + **garagem**: Number of Parking Spaces
  + **banheiro**: Number of bathrooms in the property
  + **opcionais**: List with extra features of each property (e.g., Barbecue Grill, Balcony, Orchard, etc)
  + **iptu**: Property Tax
  + **spider\_name**: Crawler name
  + **numero\_corretor**: Telefone do corretor
  + **isportal**: Boolean flag - Check if the ad is from a portal
  + **isparticular**: Boolean flag - Check if the ad is particular
  + **zoneamento**: Property Zoning
    - **id**: Zoning ID
    - **sigla**: Zoning Acronym
    - **nome**: Zoning full name
  + **area\_total**: Total area of ​​the property (in m²)
  + **tipo\_negocio**: Business Type (Sell or Rent)
  + **error**: Indicates if there was an error in the request.
  + **url**: Ad URL
  + **condominio**: Condo fee
  + **status**: Ad Status Flag (confirm)
  + **codigo**: Ad Code
  + **geometry**: Data for property location
    - **type**: How to mark on the map (dot)
    - **coordinates**: Coordinates for marking
  + **cidade**: City where the property is located
    - **full\_name**: Full name (city, state, country)
    - **short\_name**: Short name (just city)
  + **nome\_corretor**: Name of the broker
  + **cidade\_uf**: <city>\_<state>
  + **bairro**: Neighborhood
    - **id**: Neighborhood ID
    - **nome**: Neighborhood Name
  + **anunciante**: Advertiser Identifier
    - **id**: ID
    - **nome**: Full name
  + **geohash**: Geospacial locator (<https://www.movable-type.co.uk/scripts/geohash.html>)
  + **title**: Ad Title
  + **endereco**: Property Address
  + **area\_privativa**: Total Private Build Area (in m2)²

**Note 1**: Data points may not have a defined schema, that is, there may be null or unspecified fields, so it is up to you to decide how to treat them.

**Note 2**: Some attributes are complex data types (e.g., array, vectors, lists). The candidate can handle the analysis containing this attribute type as a bonus.

**Note 3**: During the process of retrieving information (crawler/spider in progress), the developer added new attributes that caused attributes with null values ​​in a specific time window.

## Minimum Requirements

* Reading of the provided dataset
* Exploratory Data Analysis
  + Visualization with dimensionality reducers
  + Principal Factors Analysis
  + Cuts to extract outliers
* Compile the information so that it is presentable for a presentation.

## Bonus items

These are things that we will be glad if you do, and they will definitely set you apart from the other candidates:

* Analysis containing complex data types(“**opcionais**” field)
* To be able to provision your entire environment in a public cloud, preferably AWS.
* Create assumptions, validate against data and prove with storytelling and graphs;

## Tools

The analysis should be done in Python or R. If you want to use another language or tool, you should contact [recrutamento@datasprints.com](mailto:recrutamento@datasprints.com) justifying why you should use it.

Cite references in case you use a more advanced tool/algorithm or framework.

## How to deliver

You must deliver to us 3(three) essential packages:

* ***README.md*** **-** file containing instructions for reproducing your analysis;
* **Analysis** - file (HTML, PDF, jupyter notebook, ipynb, R, etc.) with the answers to each topic from the [minimum requirements](#_dyfc3cx7ue8e) section and with the additional reviews from the [bonus](#_t7e1gpy2296s) section. When possible, use [storytelling techniques](https://www.columnfivemedia.com/data-storytelling-brands-data-visualization) to make your thinking clear to us;
* Source code: The codes and queries you used to build your analysis. (extra points for deliveries in repositories such as ***Github* or *Gitlab***).

The documentation should broach:

* Summary of the issue to be addressed
* A brief description of the algorithms, main functions, and procedures.
* Implementing decisions that may be missing from this document.

|  |
| --- |
| **TEST CORRECTION**  The candidate will be evaluated according to the following criteria:  Clarity  Information extracted from data  Outlier detection  Strategy decisions |