



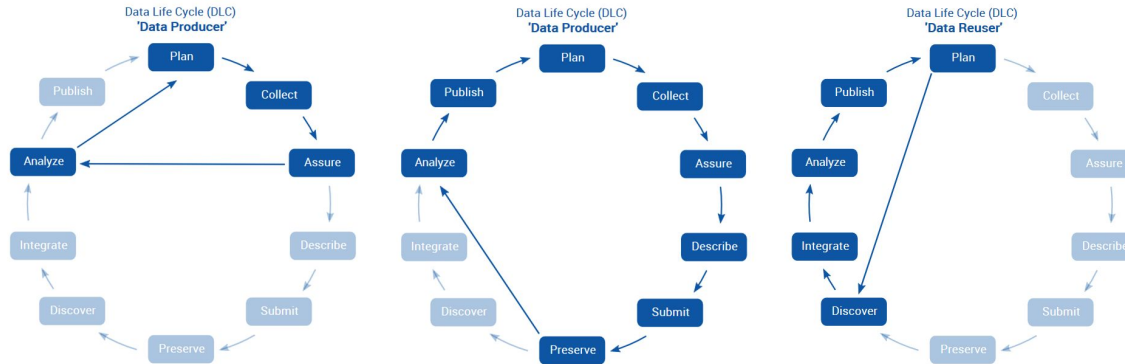
Gibt es eine Korrelation zwischen
der Positivitäts-Rate und der
Anzahl abgeschlossener Tests in
Deutschland zur Corona-Zeit?

Forschungsfrage

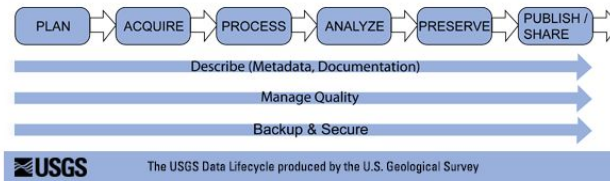
"Gibt es eine Korrelation zwischen der Positivitäts-Rate und der Anzahl abgeschlossener Tests in Deutschland zur Corona-Zeit?"

- dazu Datensatz: Archive of historical data on the testing volume for COVID-19 vom European Centre for Disease Prevention and Control
- Positivitäts-Rate = $100 \times \text{Anzahl neuer positiver Fälle} / \text{Anzahl abgeschlossener Tests pro Woche}$
 - zeigt besser, wie viele Leute wirklich krank sind
- In einer perfekten Welt:
 - es gibt genug Tests und Menschen testen sich dann, wenn sie sich krank fühlen
 - -> eine Korrelation sollte vorhanden sein
- Frage: War das in der Realität auch so?

Data Lifecycle



German Federation for
Biological Data



United States Geological
Survey

Data Lifecycle (meine Version)



Plan

- Forschungsfrage -> Anforderungen an Daten
- Data Management Plan
 - basierend auf Horizon Europe Template (auch empfohlen durch Uni Jena)
 - klärt viele Fragen bezüglich des Lifecycles
 - Datenbeschreibung
 - FAIR sicherstellen
 - Kosten
 - Datensicherheit und ethische Fragen
 - hilft dabei, nichts zu vergessen
 - "living document"
 - manche Fragen können erst im Lauf des Projektes geklärt werden
 - besonders sinnvoll in großen Projekten

Discover

- verschiedene Quellen bereits gegeben
- Datensatz von "European Centre for Disease Prevention and Control" gefunden
- Beim Suchen verschiedene Dinge überprüfen:
 - Inhalt
 - Qualität (erster Eindruck)
 - Größe des Datasets
 - Provenance
 - sensitive Daten?
 - Verfügbarkeit / Lizenz



Assure

- Qualitätsmerkmale überprüfen
 - Completeness: 100%, keine Woche fehlt, keine NaN-Einträge
 - Uniqueness: 100%, keine Wochen mehrfach im Dataset
 - Timeliness: historische Daten, kurz nach Ende des Erfassungszeitraums veröffentlicht
 - Validity: alle Datenpunkte haben korrekten Typ nach Import als Dataframe
 - Accuracy: schwer zu sagen, einfache plots sehen sinnvoll aus, Quelle ist verlässlich
 - Consistency: nur ein Datensatz genutzt -> dieser ist konsistent
- einige Preprocessing Schritte bereits benötigt zur Überprüfung

Process

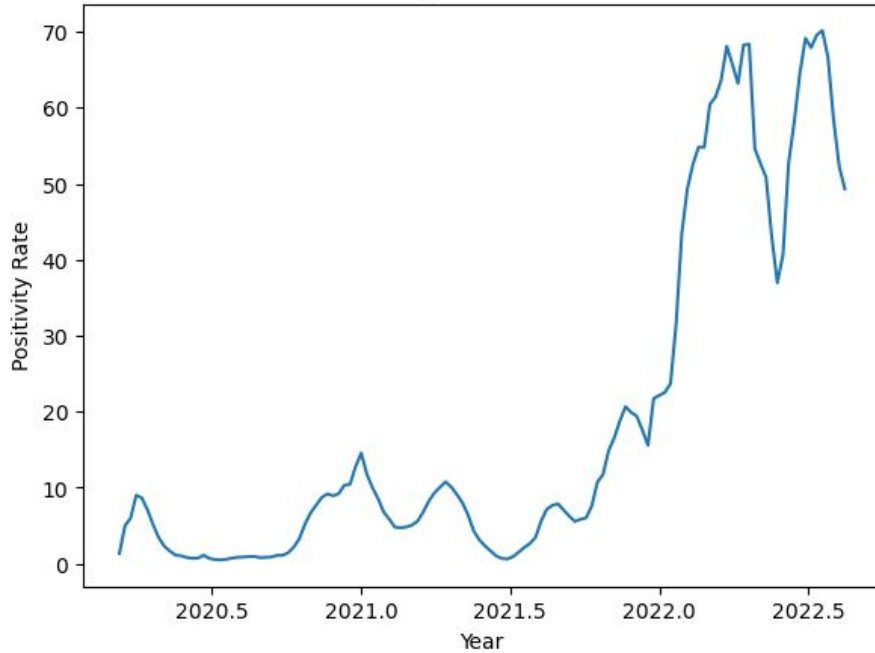
- Filtern der Daten in Deutschland auf level "national"
- Herausfiltern von unnötigen Spalten
- Herausfiltern von NaN-values
- Spalte year_week: string("2020-W35")
 - unpraktisch für statistische Analyse
 - daher Aufsplitten in 3 Spalten:
 - year
 - week
 - time
 - Spalte time = year + week/53
 - ermöglicht vergleichbare Daten (z.B. 2020.04 < 2020.36)
- Erstellen einer neuen Datei mit vorverarbeiteten Daten

Describe

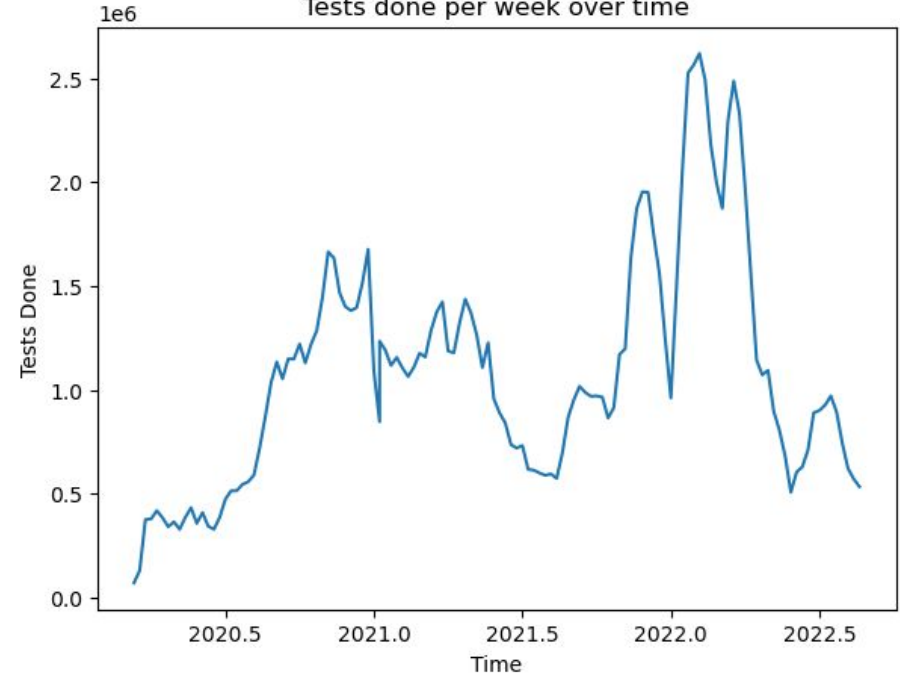
- Beschreibung der Verarbeitungsschritte
 - festgehalten in jupyter notebook
 - markdown + code
 - somit sogar reproduzierbar
 - dafür nötig: Abhängigkeiten beschreiben
 - Programmiersprache, Packages + Versionen, Installationsanleitung
- Beschreiben der Daten -> Metadaten
 - festgehalten in markdown-file und xml file
 - somit human- und machine-readable
 - Inhalt:
 - Kurzbeschreibung, Autor, Spalten, Zeitraum, Typ, Veröffentlichungsdatum, Identifier, Contributor, Nutzungsrechte, ...

Analyze

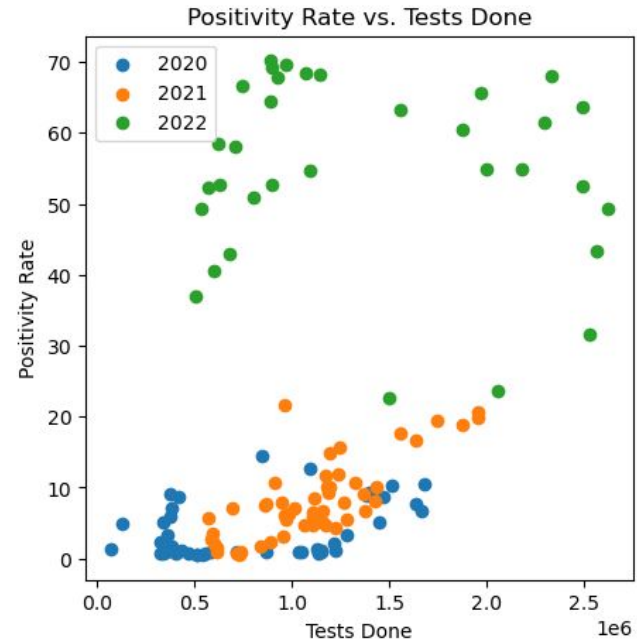
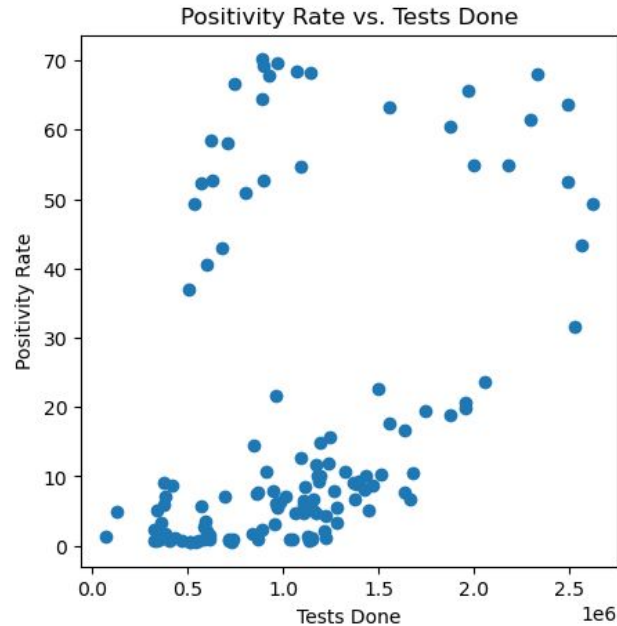
Positivity Rate over time



Tests done per week over time



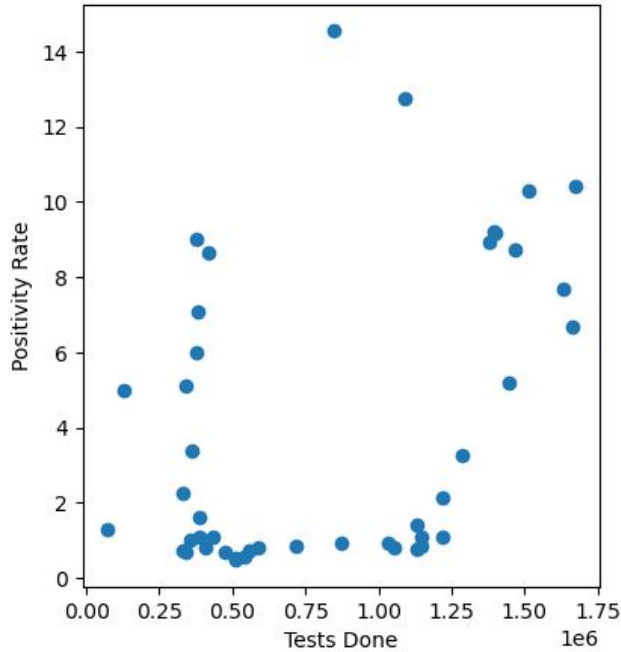
Analyze



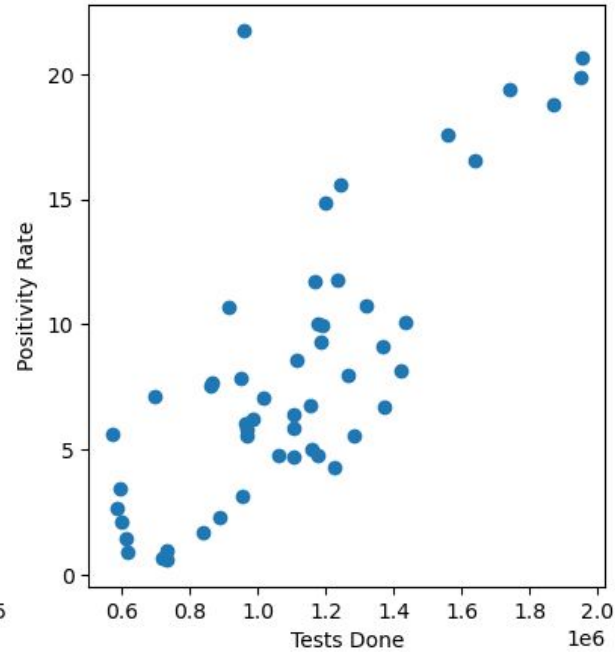
Analyze

Positivity Rate vs. Tests Done, red = week 1, yellow = week 53

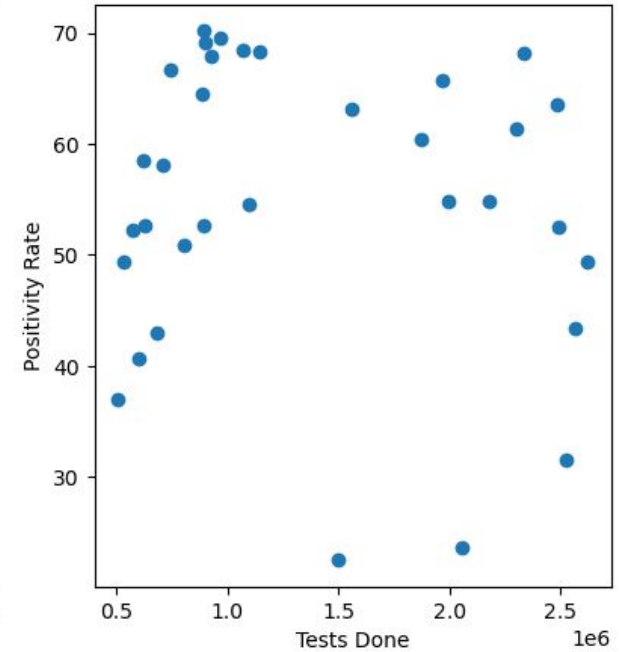
2020, Correlation: 0.41



2021, Correlation: 0.77



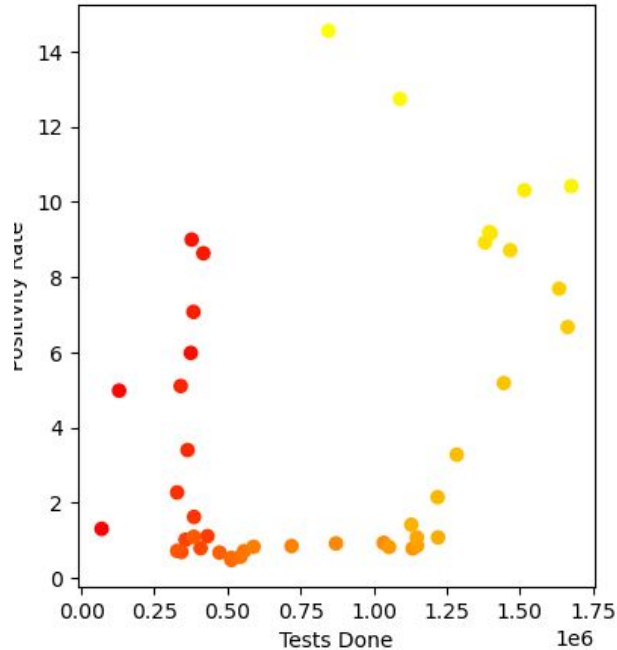
2022, Correlation: -0.13



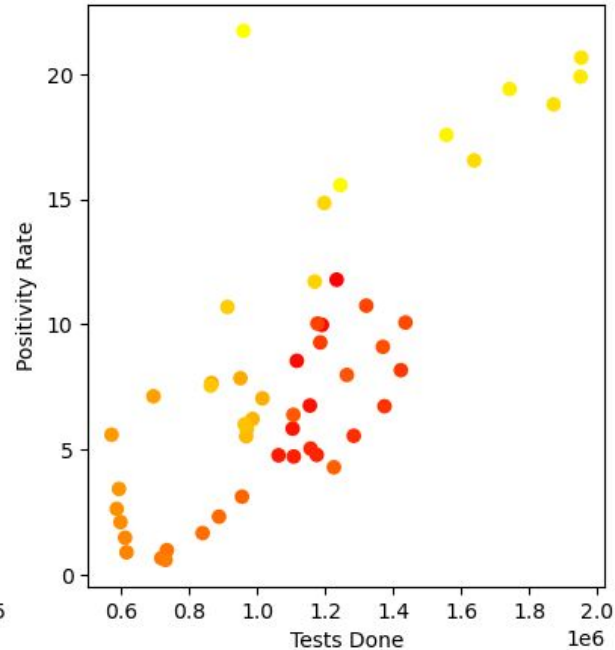
Analyze

Positivity Rate vs. Tests Done, red = week 1, yellow = week 53

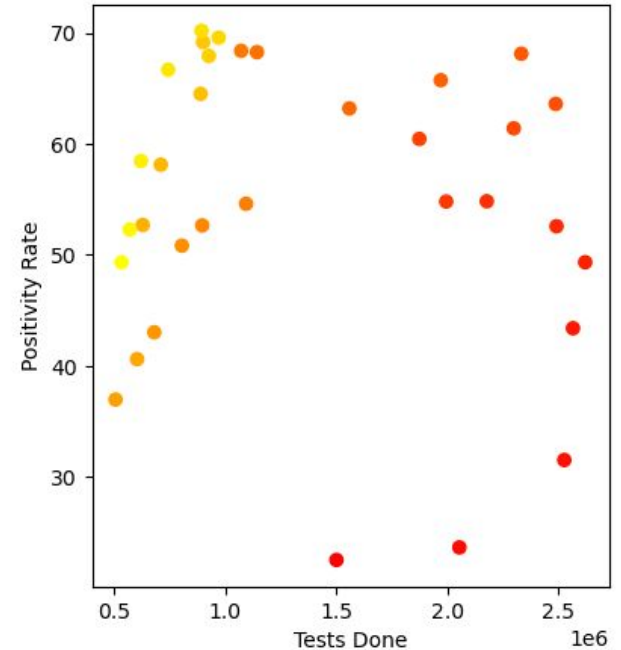
2020, Correlation: 0.41



2021, Correlation: 0.77



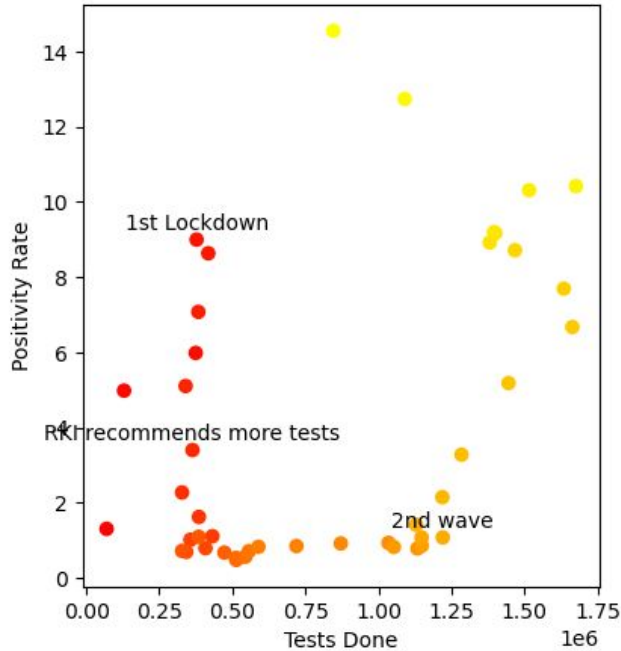
2022, Correlation: -0.13



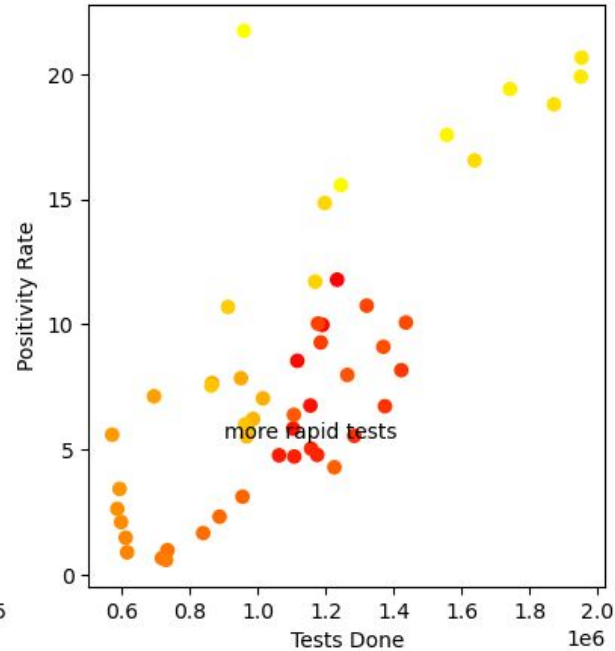
Analyze

Positivity Rate vs. Tests Done, red = week 1, yellow = week 53

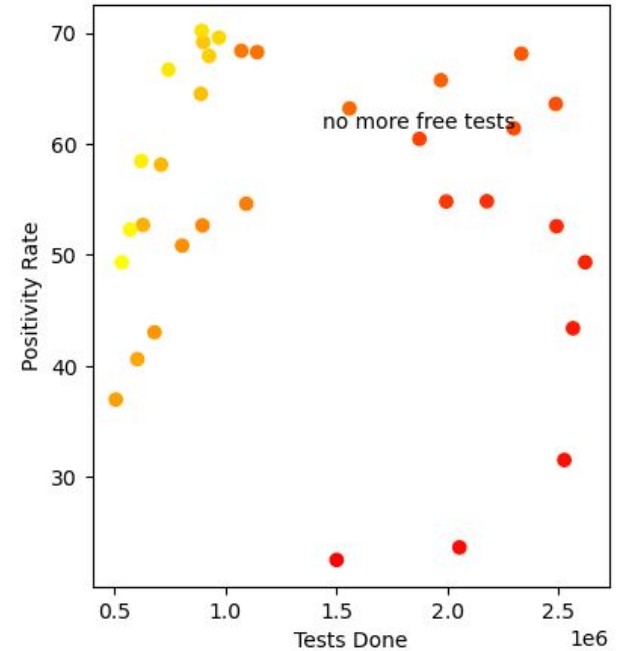
2020, Correlation: 0.41



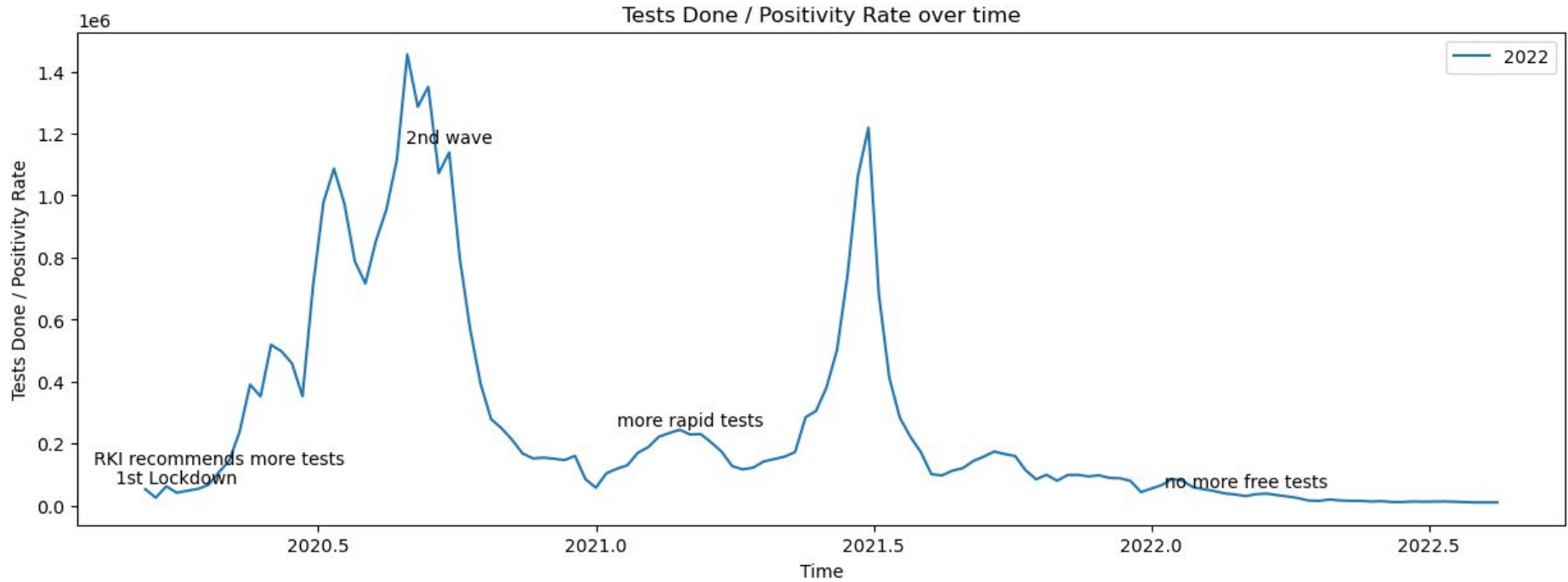
2021, Correlation: 0.77



2022, Correlation: -0.13



Analyze



Analyze

"Gibt es eine Korrelation zwischen der Positivitäts-Rate und der Anzahl abgeschlossener Tests in Deutschland zur Corona-Zeit?"

- Korrelation besteht in manchen Zeiträumen
- besonders Ende 2020 und 2021 das ganze Jahr über
- in anderen Zeiträumen kaum vorhanden
 - Gründe: fehlende Tests, keine kostenlosen Tests
- verschiedene Ereignisse / politische Beschlüsse spiegeln sich in Daten wieder

Publish and Preserve

- Ziel: Erfüllung der FAIR-Kriterien
- **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
- Findable:
 - alles hochgeladen in öffentlichem GitHub Repository
 - GitHub sollte im Index verschiedener Suchmaschinen erfasst sein
 - zusätzlich möglich: Anfragen nach Indexing
 - alle Dateien im Repo haben einzigartige Identifier
 - maschinenlesbare Metadaten sind vorhanden
 - enthalten wichtige keywords
 - Synchronisation mit Research Data Repository (Zenodo)

Publish and Preserve

- Accessible:
 - GitHub ist vertrauenswürdig + sehr verbreitet
 - Herunterladen:
 - automatisch: git + ssh / https, GitHub for command line
 - manuell von Website
 - Identifier auf Zenodo sind persistent
- Interoperable:
 - Daten als csv abgespeichert
 - sehr verbreitet, von vielen Tools unterstützt
 - Spaltennamen sind selbsterklärend und oft verwendet
 - maschinenlesbare Metadaten erstellt mit "dublin core generator"
 - -> verwenden somit typisches Vokabular

Publish and Preserve

- Reusable
 - Nutzung von jupyter notebooks für Preprocessing
 - reproduzierbar
 - nachvollziehbar
 - Abhängigkeiten + Installation beschrieben
 - MIT non-copyleft license für Code, CC0 license für Daten
 - erlaubt freie Wiederverwendung
 - Datenqualität ist hoch
 - Datensatz wird sich nicht mehr ändern
 - -> Repository wird archiviert



—
Vielen Dank für ihre
Aufmerksamkeit!